Predicción de Morosidad

Marié del Valle Reyes*
*Facultad de Matemáticas y Ciencias de la Computación,
Universidad de La Habana, La Habana, Cuba

1. Problema

Una entidad bancaria está interesada en identificar comportamientos de clientes que puedan derivar en morosidad. Se tiene la información histórica de clientes (*ID*) a los que les fueron aceptadas sus solicitudes de crédito, conociéndose si el clientes incurrió (variable *default* = 1) o no (variable *default* = 0) en mora. Las posibles variables de entrada a utilizar para identificar patrones de morosidad son las siguientes:

- *Age*: Edad.
- *Income*: Ingresos anuales.
- *Exp_Inc*: Proporción del gasto mensual de la tarjeta de crédito para ingresos anuales.
- *Avgexp*: Gasto mensual de la tarjeta de crédito.
- *Ownrent*: Si es propietario de una vivienda.
- *Selfempl*: Si es trabajador por cuenta ajena.
- *Depndt*: 1 + Número de dependientes.
- *Inc_per*: Ingresos divididos entre el número de dependientes.
- *Cur_add*: Meses viviendo en la misma dirección.
- *Major*: Número de tarjetas de crédito a cargo.
- *Active*: Número de cuentas activas de crédito.

Teniendo en cuenta lo anterior, construye un modelo que permita predecir la morosidad de los clientes.

2. Análisis Exploratorio de Datos

La tabla 1 proporciona un resumen de las variables del problema, incluyendo su tipo, estadísticas descriptivas como mínimo, máximo, media, mediana y desviación típica.

Las variables Ownrent, Selfempl y default son binarias. La variable objetivo es default, la cual tiene una proporción de 1's de 10.47 %, lo cual significa que aproximadamente el 10.47 % de las observaciones están etiquetadas como clientes que incurren en mora, mientras que el 89.53 % restante están etiquetadas como cliente que no lo hacen. Además, es importante destacar que ninguna de las variables en el conjunto de datos presenta valores faltantes o missings, lo que garantiza la integridad y completitud de los datos para el análisis subsiguiente.

2.1. Correlación de las variables

En la figura 1, se muestra una matriz de correlación entre las variables continuas del problema. Las variables cuya correlación es mayor respecto a las demás son Exp_Inc y Avgexp con un valor de 0.84. Esto se debe a que ambas variables estan relacionadas con el gasto mensual de la tarjeta de crédito.

Cuadro 1: Resumen de Variables Entradas y Variable Objetivo (default).

Variable	Tipo de variable	Mínimo	Máximo	Media	Mediana	Desviación típica
Age	numeric	0.1667	83.5	33.17	30.67	10.42
Income	numeric	1.5	12	3.44	3	1.72
Exp_Inc	numeric	0.0001859	0.9063	0.09	0.061	0.11
Avgexp	numeric	0	2291.17	241.76	156.09	291.09
Ownrent	integer	0	1	0.49	0	0.50
Selfempl	integer	0	1	0.06	0	0.24
Depndt	integer	0	6	0.96	0	1.24
Inc_per	numeric	0.45	11	2.20	2	1.33
Cur_add	integer	0	540	53.99	28	63.73
Major	integer	0	1	0.84	1	0.37
Active	integer	0	29	7.35	6	6.13
default	integer	0	1	0.10	0	0.31

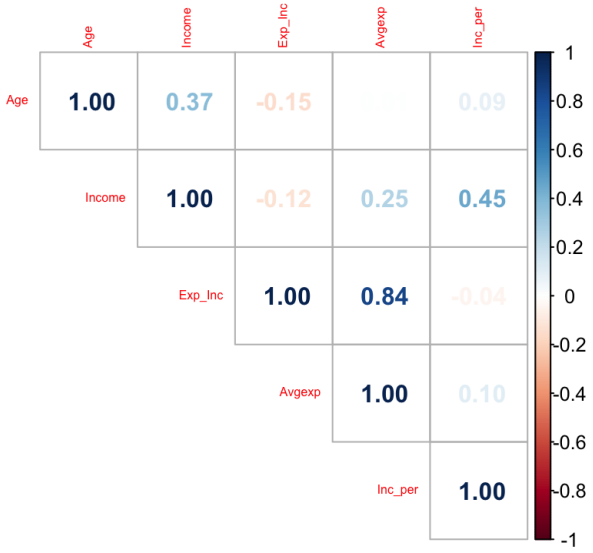

Figura 1: Matriz de correlación entre variables continuas.

3. Manipulación de los datos

El hecho de que el valor mínimo de la variable Age sea 0.1667 sugiere la presencia de valores atípicos (outliers) en los datos. Estos valores no son consistentes con el contexto del problema. El siguiente gráfico (Fig 2) muestra la presencia de valores atípicos en los datos de la variable Age.

Para abordar el problema de los valores atípicos, hay dos enfoques comunes que se pueden considerar. El primero implica eliminar los valores que se encuentran por debajo o por encima de ciertos percentiles de los datos. La Figura 3 muestra los datos de la variable Age después de eliminar los datos cuyos cuartiles están por debajo del 0.5 % o por encima del 99.8 %.

La segunda opción implica transformar los datos aplicando el logaritmo a la variable. Al aplicar el logaritmo, los valores extremadamente grandes se reducen y los valores extremadamente pequeños se amplifican, lo que puede ayudar a normalizar la distribución de los datos y reducir el impacto de los valores atípicos en el análisis. La siguiente figura (Fig. 3) muestra los histogramas de la variable Age antes y después de aplicar el logaritmo a la variable.

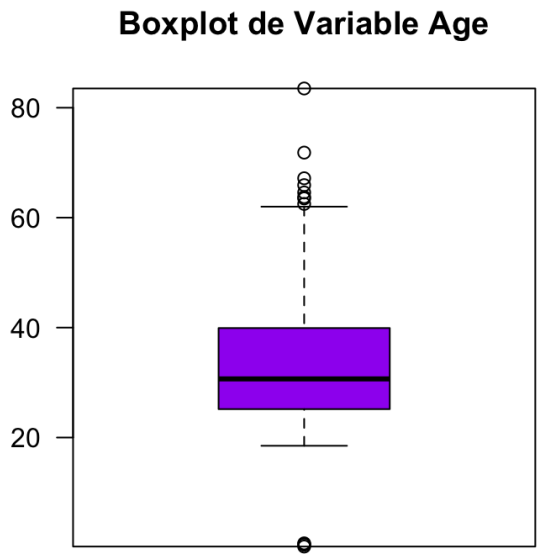

Figura 2: Boxplot de los datos de la variable Age, en el que se aprecian valores atípicos.

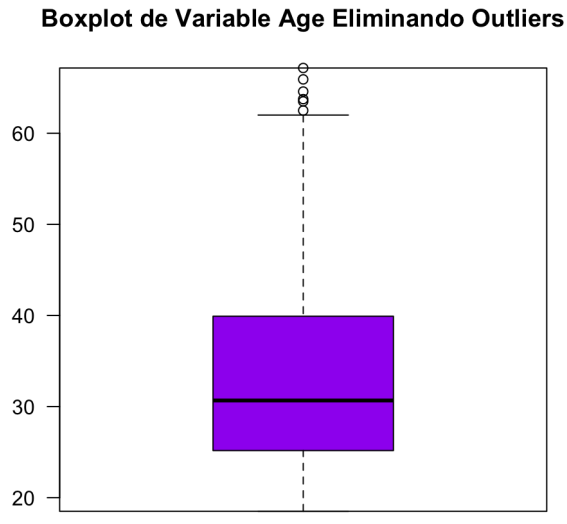

Figura 3: Boxplot de los datos de la variable Age. Se eliminaron datos por debajo del percentil 0.05 y por encima del percentil 0.998.

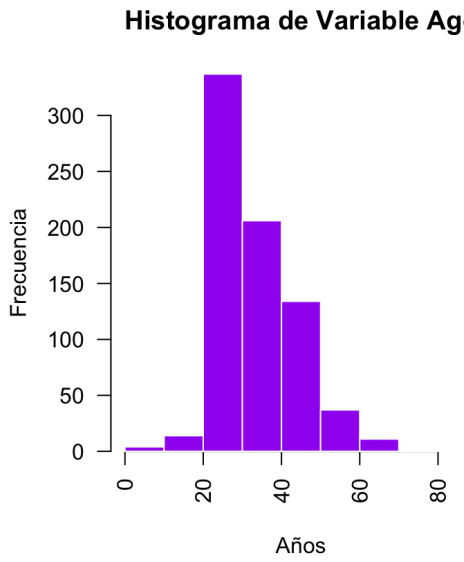

Figura 4: Histograma de la variable Age.

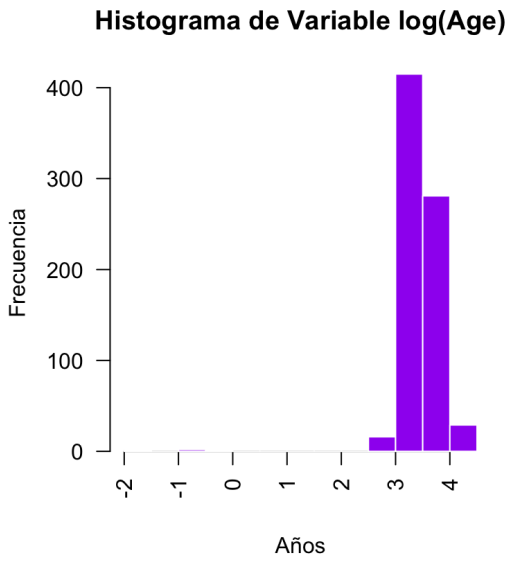

Figura 5: Histograma de la variable Age después de aplicarle el logaritmo.

4. Modelo

En este estudio, se llevó a cabo el análisis de un modelos de aprendizaje automático: regresión logística, con el objetivo de predecir la probabilidad de morosidad en un conjunto de datos financiero. Para garantizar la integridad y la calidad de los datos, se utilizó el conjunto de datos de entrenamiento (train), en el cual se eliminaron los valores extremos considerados atípicos. Específicamente, se excluyeron los datos cuyos percentiles eran mayores que el percentil 99.8 y menores que el percentil 0.5.

4.1. Regresión logística

La regresión logística se utiliza para predecir la probabilidad de que ocurra un evento binario en función de una o más variables predictoras.

Primero, se realizó el proceso de selección de variables, en el cual se utilizó el método de pasos. Dicho método consiste en encontrar el modelo óptimo que mejor se ajuste a los datos al considerar diferentes combinaciones de variables explicativas. Se comienza con un modelo simple que solo incluye el término de intercepción y se avanza hacia un modelo más complejo que incluye todas las variables disponibles. Todos los modelos intermedios se evalúan y comparan utilizando el criterio de información de Akaike (AIC). Finalmente se selecciona el modelo con el valor de AIC más bajo como el mejor modelo, equilibrando así la precisión predictiva con la complejidad del modelo.

El modelo resultante fue el siguiente:

$$default = Active + Avgexp + Inc\_per + Cur\_add \tag{1}$$

Variable	Estimate	Std. Error	Valor z	Pr(>|z|)
(Intercept)	-3.7152	0.3379	-10.9936	< 2.2e-16
Active	0.0763	0.0184	4.1483	3.35e-05
Avgexp	0.0010	0.0003	2.9190	0.00351
Inc_per	0.1869	0.0810	2.3064	0.0211
Cur_add	0.0036	0.0016	2.2132	0.0269

Cuadro 2: Estimaciones de los coeficientes de la regresión logística

En la tabla 2 se observa que los p-valores que indican el nivel de significancia estadística de cada

coeficiente, siendo la variable Active la que mayor valor tiene, lo cual prueba su importancia en la predicción de morosidad. Las variables Inc_per y Cur_add son moderadamente significativas.

5. Evaluación

Al aplicar el modelo al conjunto de datos test, se contruyó una curva ROC (Fig. 6) y se calculó el AUC. El valor de AUC es 0.7976889, lo cual indica que el modelo tiene una buena capacidad para clasificar correctamente los casos positivos y negativos en el problema de clasificación binaria.

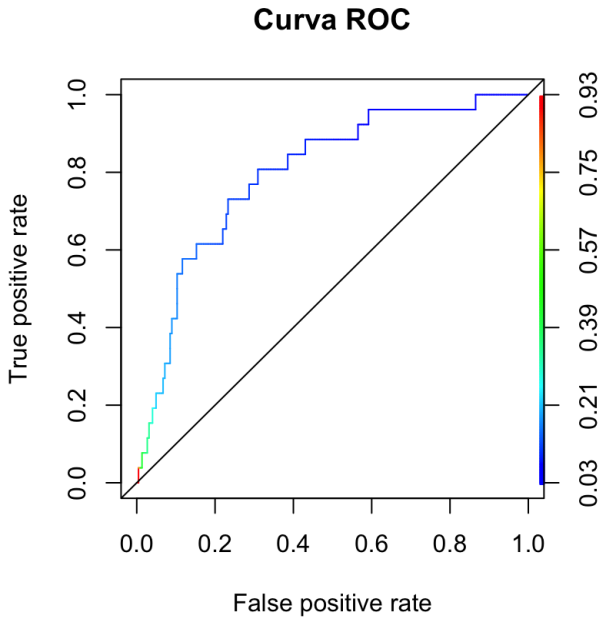

Figura 6: Curva ROC para modelo de regresión logística.

6. Repositorio Github

Los datos, código y docuemntación se encuentran en el repositorio de github default-prediction.