Predicción de clientes con mal pago

Por: Juan Sebastián Acevedo R
Para: Data Team - BBVA

Análisis descriptivo

El análisis descriptivo se llevó a cabo en dos fases.

1. Análisis Exploratorio Univariado:

Observe el reporte interactivo en el archivo `1.1_analisis_descriptivo_interactivo.html`
El código para crear el reporte se encuentra en `1.1_analisis_descriptivo_interactivo.ipynb`

Para cada variable se presenta:

- Medidas estadísticas básicas (media, moda, desviación)
- Distribución y gráfico de frecuencias
- Conteo de individuos de cada categoría para variables categóricas
- Límites superior e inferior para variable continuas
- Valores faltantes
- Alertas importantes sobre desbalance de clases.

Principales Hallazgos:

- ❖ La cédula es la llave única para registro, no existen registros repetidos, los datos con consistentes en su mayoría.
- ❖ Existe un desbalance de clase para la variable objetivo, únicamente el 2% de clientes está tipificado como mal pago.
- ❖ Las únicas columnas con valores faltantes son tipo_vivienda y tipo_contrato
- ❖ Las cantidades monetarias se encuentran en montos viables, no obstante, se observa que de acuerdo a la regla intercuantil, hay valores atípicos. (Pero no inconsistentes)

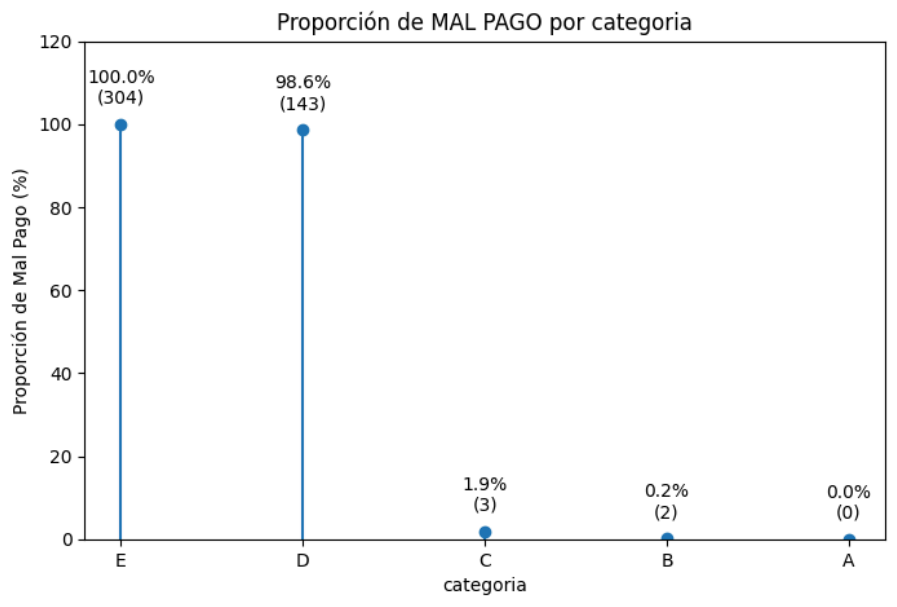
2. Análisis Descriptivo Bivariado:

Observe el notebook `1.2_analisis_descriptivo_multivariado.ipynb`
Para entender visualmente la relación de cada variable explictiva con la variable objetivo se tomaron principalmente dos métodos:

- ★ BoxPlots de cada variable cuantitativa diferenciando por clase de cliente (bueno-malo)
- ★ Para variables categóricas, Lollipops representando la proporción de malos clientes en cada segmento.

Pricipales Hallazgos:

- ❖ La variable CATEGORIA, cuyos valores son A, B, C, D, E separa casi perfectamente los clientes buenos y malos.



Vemos que todos los malos clientes se concentran en las categorías D y E, esta clasificación es tan perfecta que sospechamos que no corresponde a ninguna característica “transaccional” de los clientes, sino que hace parte de la respuesta, por lo tanto, no la vamos a **considerar como explicativa.**

- Para el caso de la variable Dias de Mora, está estrechamente relacionada con la variable objetivo, no debería considerarse como una variable explicativa, ya que la idea es identificar a los clientes mala paga antes de que empiecen a tener mora en sus pagos.

- Se hace un análisis de regresión preliminar para confirmar algunas conclusiones hechas a partir de la visualización.

Algoritmo Probabilístico

Para observar el código relacionado con esta sección diríjase a `1.2_desarrollo_modelo.ipynb`

Construimos un algoritmo probabilístico basado en un modelo XGBoost de Machine Learning. Para ello, fue necesario **codificar** las variable categoricas no numéricas, posteriormente entrenar el modelo y realizar calibración de hiperparámetros mediante el método V**alidación Cruzada con estratificación**, para abordar el problema de **desbalance de clase.**

Codificación de variables:

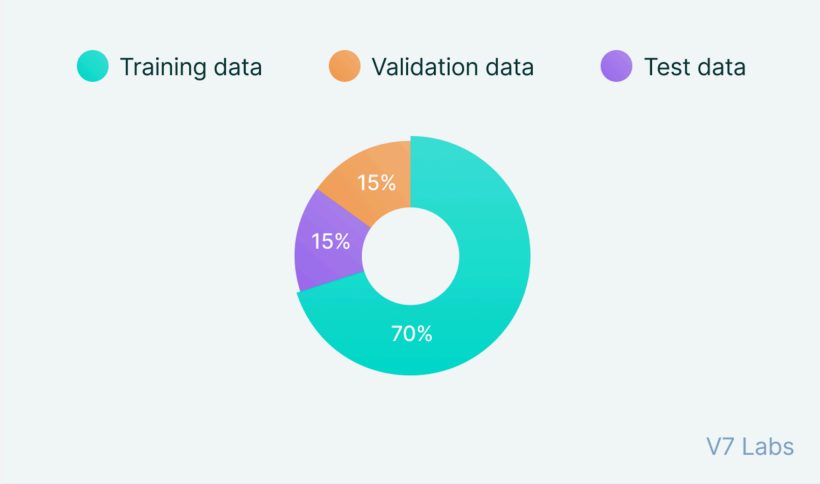
- Las variables **nominales**, que no poseen un orden natural, se codificaron usando variables **dummy**. Las cuales son : 'oficina', 'ocupacion', 'estado_civil' y 'sexo'.
- Para las variables enteras como edad, personas_a_cargo y num_creditos, se discretizaron usando deciles o creando grupos teniendo en cuenta la distribución.

- ➔ Las variables ordinales, en las que subyace un orden natural, se codificaron usando la técnica **ordinal encoder,** en el caso de presentarse valores faltantes se asignó un entero negativo, simulando la existencia de otra categoría (faltante). Las cuales son 'edad_deciles', 'nivel_educativo', 'personas_discreta', 'num_creditos_discreta' y 'antiguedad_entidad_discreta'.
- ➔ Las variables numéricas contínuas no se transformaron

Partición Estratificada

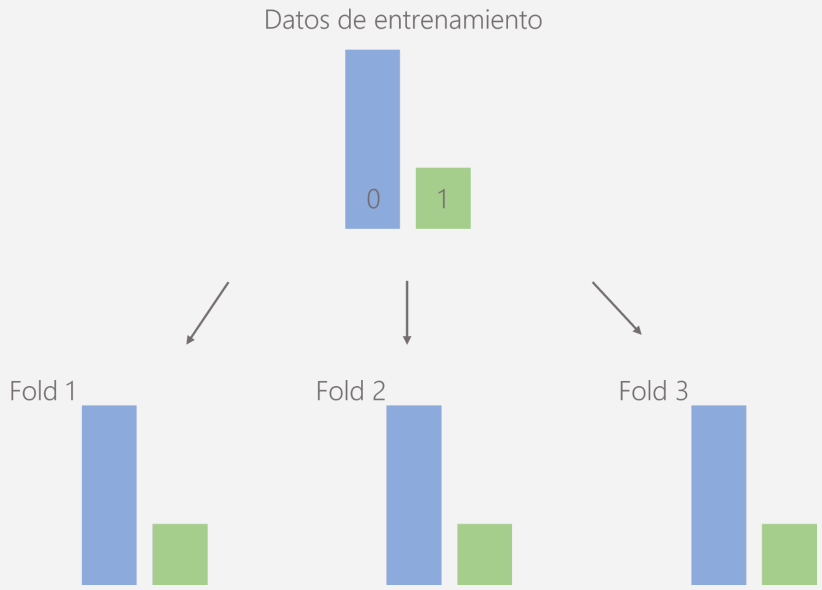
Una técnica para manejar el **desbalance** de clases en problemas de clasificación supervisada, consiste en garantizar que todos los subconjuntos y grupos de datos tengan la misma proporción de desbalance de clase. De esta forma, en principio no es necesario realizar **oversampling ni undersampling**.

Para construir el modelo predictivo y reportar su desempeño usaremos particiones de la forma Train - Validation - Test. En donde los conjuntos tendrán los siguientes tamaños:



Debido al desbalance de clases, tanto el conjunto de test como el de train deben tener la misma proporción de desbalance de clase, que en nuestro caso es 0.2.

Adicionalmente, para evitar el sobreajuste y tener buena capacidad de generalización, en la calibración de hiperparámetros se implementa la técnica **Stratified Kfold Cross Validation,** que consiste en la partición de los conjuntos train y validation en grupos o folds que preservan la proporción de desbalance de clases.



EL XGBoost es un algoritmo de ensamble basado en árboles de decisión, es un algoritmo del estado del arte que han demostrado excelente desempeño en tareas de clasificación, además, dada su naturaleza de árobl de decisión, se trata de un modelo interpretable. Probamos las diferentes ténicas para controlar el desbalance de clases mediante un XGboost cuyos hiperparámetros son ajustados usando GridSearch.

Función de Pérdida / Métrica del algoritmo

Puesto que el principal objetivo es identificar a los malos clientes para tomar acciones **prescriptivas**, usamos la métrica **recall,** que consiste en calcular la cantidad de individuos que el algoritmo logra predecir como malos y dividirlo en el total de individuos que realmente cayeron en mora. En otras palabras, si tomamos a los positivos como clientes que cayeron en mora,

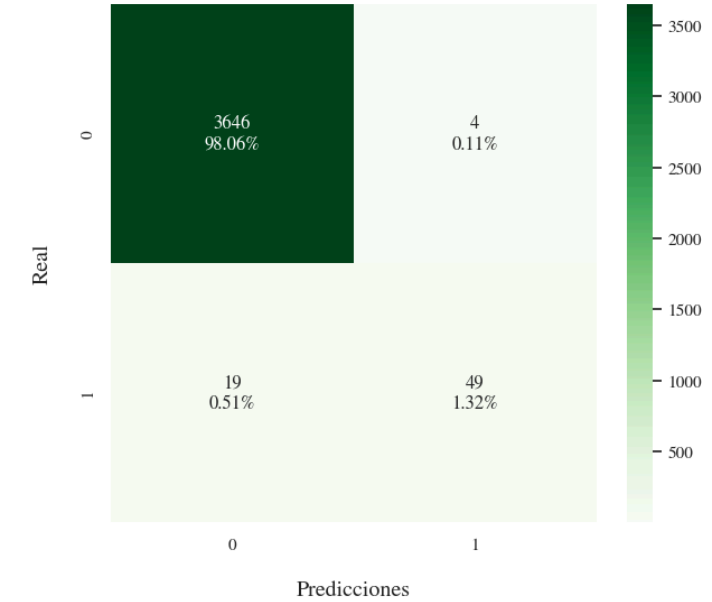
recall = reales_positivos / (reales_positivos + falso_negativo)

De esta forma, priorizamos que el modelo tal vez prediga más impagos de los reales.

Resultados

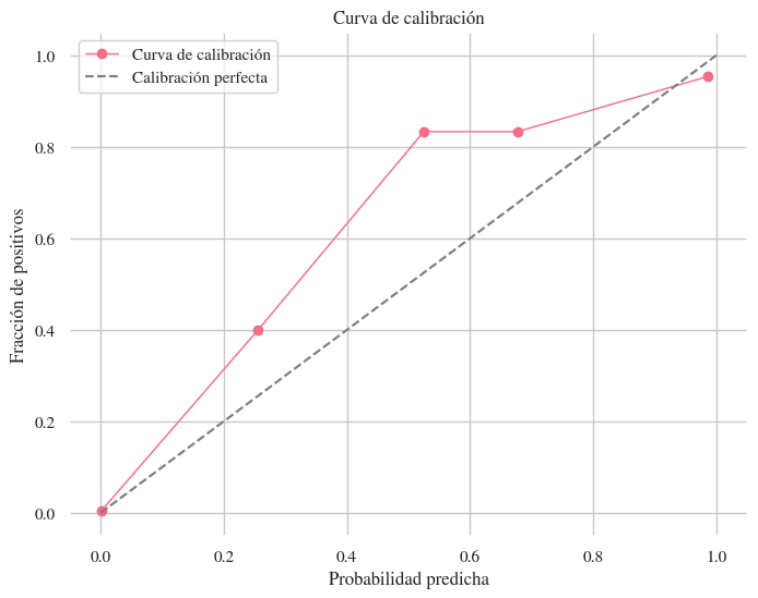
El algoritmo seleccionado tiene una precisión estimada de 99,4% en general, y una sensibilidad de 72% para identificar clientes que caerán en mora.
A continuación se observa la matriz de confusión del modelo en “data nueva”

F1 Score	0.810
Accuracy	0.994
Recall	0.721
Precision	0.925
ROC-AUC	0.860



En un total de 3718 clientes, realmente 68 (1.8%) caerían en mora, y el algoritmo sería capaz de identificar a 49 (1.32%) de ellos.

Adicionalmente, el algoritmo nos brinda la **probabilidad** de que un cliente caiga en mora. Mediante la siguiente **curva de calibración**:



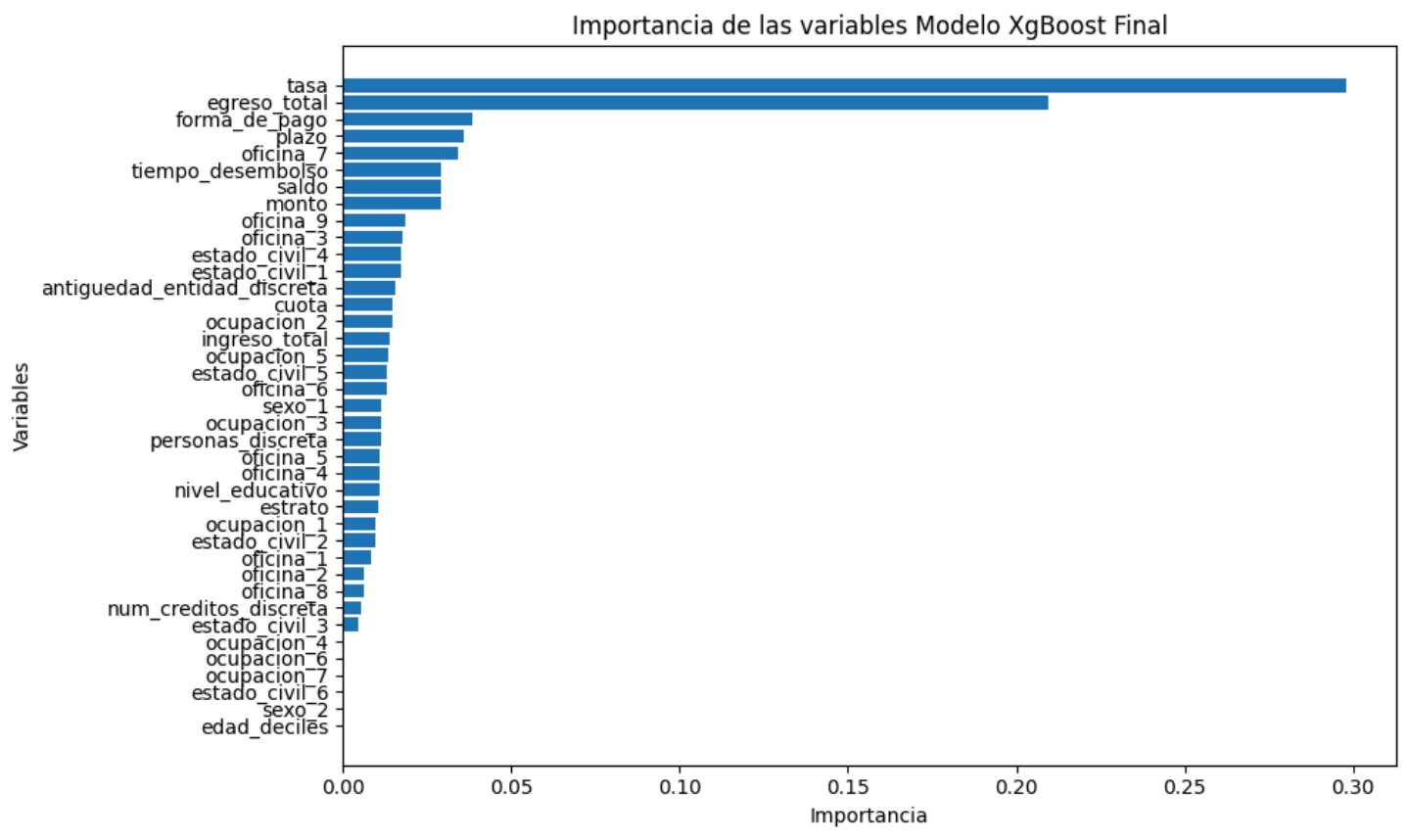
Observamos que

- Cuando el modelo realiza predicciones con probabilidad cercana a 0, es casi seguro que el cliente será bueno.

- Cuando el algoritmo realiza predicciones con probabilidad cercana a 1, es casi seguro que el cliente será malo
- Cuando la probabilidad está cercana a 0.5, el algoritmo prefiere clasificarlo como mal pago, de esta manera nos aseguramos que el umbral sea suficientemente flexible.

Interpretación del Algoritmo

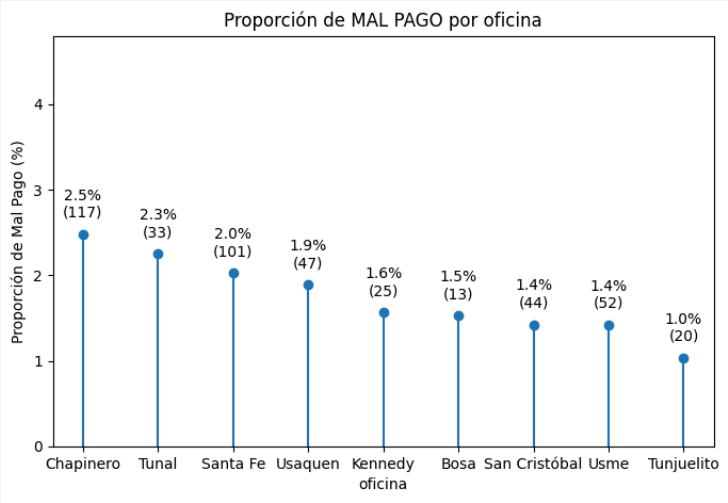
El algoritmo le otorga la importancia a las variables de la siguiente forma:



Observamos que:

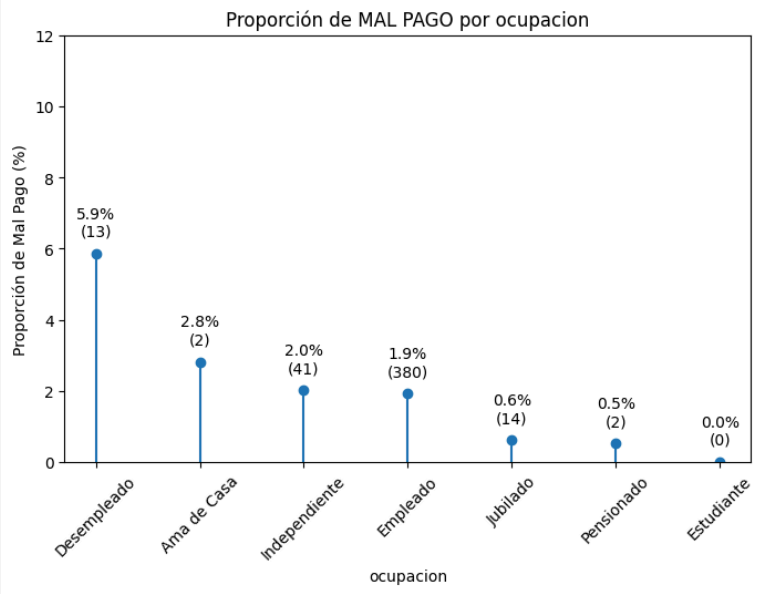
- ❖ La TASA es la variable más decisiva.
- ❖ El egreso total es muy importante.
- ❖ La edad, el sexo, el número de créditos, y el número de personas a cargo son prácticamente irrelevantes.
- ❖ Algunas oficinas tienen mayor probabilidad de otorgar créditos a malos clientes.
- ❖ Algunos segmentos de estado civil y ocupación son importantes, y otros irrelevantes.

Oficinas:



Las oficinas más críticas son las de Chapinero, Tunal y Santa Fe.

Ocupaciones:



Los desempleados tienden a caer más fácilmente en mora.

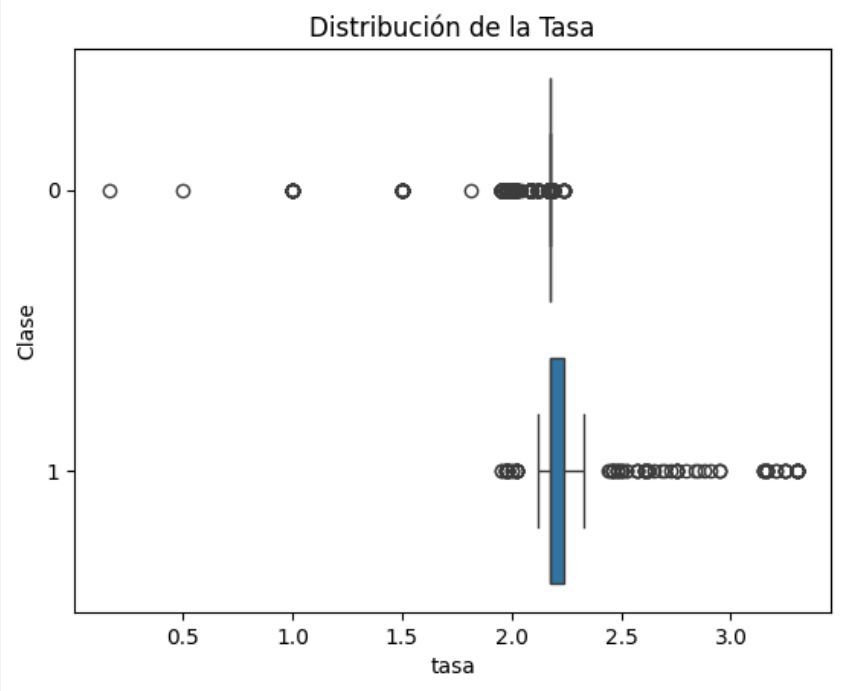
Para entender la relación entre las variables más importantes y la objetivo, realizamos un **Análisis de Regresión, también conocido como ANOVA.**

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                  clase   No. Observations:                24786
Model:                          Logit   Df Residuals:                    24749
Method:                           MLE   Df Model:                           36
Date:                Tue, 12 Mar 2024   Pseudo R-squ.:                  0.6566
Time:                        12:25:31   Log-Likelihood:                -775.43
converged:                      False   LL-Null:                       -2257.8
Covariance Type:            nonrobust   LLR p-value:                     0.000
==================================================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
oficina_1                     -1.8060   2.04e+06  -8.86e-07      1.000      -4e+06       4e+06
oficina_2                     -1.5278   2.04e+06   -7.5e-07      1.000   -3.99e+06    3.99e+06
oficina_3                     -1.6681   2.04e+06  -8.19e-07      1.000   -3.99e+06    3.99e+06
```

oficina_4	-1.9159	2.03e+06	-9.42e-07	1.000	-3.99e+06	3.99e+06
oficina_5	-1.9396	2.04e+06	-9.52e-07	1.000	-3.99e+06	3.99e+06
oficina_6	-1.5158	2.04e+06	-7.44e-07	1.000	-4e+06	4e+06
oficina_7	-1.7039	2.04e+06	-8.36e-07	1.000	-3.99e+06	3.99e+06
oficina_8	-1.8517	2.04e+06	-9.09e-07	1.000	-3.99e+06	3.99e+06
oficina_9	-1.0979	2.04e+06	-5.38e-07	1.000	-4e+06	4e+06
ocupacion_1	-0.0412	1.4e+06	-2.94e-08	1.000	-2.74e+06	2.74e+06
ocupacion_2	0.1988	1.4e+06	1.42e-07	1.000	-2.74e+06	2.74e+06
ocupacion_3	0.3658	1.4e+06	2.61e-07	1.000	-2.75e+06	2.75e+06
ocupacion_4	0.9364	1.4e+06	6.69e-07	1.000	-2.74e+06	2.74e+06
ocupacion_5	-0.9155	1.4e+06	-6.54e-07	1.000	-2.74e+06	2.74e+06
ocupacion_6	-0.6120	1.4e+06	-4.37e-07	1.000	-2.75e+06	2.75e+06
ocupacion_7	-14.9589	1.4e+06	-1.07e-05	1.000	-2.75e+06	2.75e+06
estado_civil_1	-1.1121	2.44e+06	-4.56e-07	1.000	-4.78e+06	4.78e+06
estado_civil_2	-0.7737	2.44e+06	-3.17e-07	1.000	-4.79e+06	4.79e+06
estado_civil_3	-1.1149	2.44e+06	-4.57e-07	1.000	-4.78e+06	4.78e+06
estado_civil_4	-0.1158	2.44e+06	-4.74e-08	1.000	-4.79e+06	4.79e+06
estado_civil_5	-0.0967	2.44e+06	-3.97e-08	1.000	-4.77e+06	4.77e+06
estado_civil_6	-11.8135	2.44e+06	-4.84e-06	1.000	-4.78e+06	4.78e+06
sexo_1	-7.2121	2.06e+06	-3.51e-06	1.000	-4.03e+06	4.03e+06
sexo_2	-7.8146	2.06e+06	-3.8e-06	1.000	-4.03e+06	4.03e+06
edad_deciles	15.0267	nan	nan	nan	nan	nan
nivel_educativo	0.0625	0.045	1.391	0.164	-0.026	0.151
personas_discreta	-0.1532	0.083	-1.841	0.066	-0.316	0.010
num_creditos_discreta	0.1364	0.059	2.303	0.021	0.020	0.253
antiguedad_entidad_discreta	-0.2051	0.076	-2.686	0.007	-0.355	-0.055
tiempo_desembolso	0.6233	0.027	23.511	0.000	0.571	0.675
plazo	-0.2424	0.014	-17.394	0.000	-0.270	-0.215
forma_de_pago	-0.7031	0.453	-1.551	0.121	-1.592	0.186
estrato	-0.0042	0.107	-0.040	0.968	-0.214	0.206
monto	-4.935e-06	2.59e-07	-19.038	0.000	-5.44e-06	-4.43e-06
saldo	4.933e-06	2.39e-07	20.680	0.000	4.47e-06	5.4e-06
tasa	9.7359	4.008	2.429	0.015	1.880	17.592
cuota	2.142e-05	2.5e-06	8.567	0.000	1.65e-05	2.63e-05
ingreso_total	5.255e-07	1.55e-07	3.399	0.001	2.22e-07	8.29e-07
egreso_total	-5.139e-06	5.64e-07	-9.108	0.000	-6.25e-06	-4.03e-06

- Entre mayor sea el saldo, es más probable el impago.
- Entre más grande sea la cuota y la tasa, es más probable el impago.

AQUELLOS CON UNA TASA SUPERIOR al 2.17% tienden a caer en Mora.



Recomendaciones

- ➢ Buscar planes de **acuerdo** para disminuir la **tasa** de los clientes que tienen una mayor al 2.17%
- ➢ Realizar **seguimiento** a los clientes con **egresos** muy grandes y recomendarles planes de ahorro.
- ➢ Realizar **seguimiento** a clientes con **saldos** muy altos y asegurar esos préstamos con pólizas que resguarden al banco.
- ➢ Definir una **meta agresiva** de indicador de impagos para la oficina de **Chapinero**
- ➢ Promover seguros de **desempleo**
- ➢ Incentivar el método de pago 2
- ➢ Incentivar préstamos a personas con nivel de escolaridad superior a Bachillerato.

- ★ Desplegar un modelo predictivo recurrente todas las semanas para identificar a clientes con alto riesgo de impago y priorizarlos.