# Informe de identificación predeterminada del cliente que aborda

El contenido del conjunto de datos consiste, en 30.000 registros de información sobre pagos predeterminados, factores demográficos, datos crediticios, historial de pagos y estados de cuenta de clientes de tarjetas de crédito desde abril de 2005 hasta septiembre de 2005.

Hay 25 variables, las cuales se muestran en detalle, a continuación:

* **ID**: identificador de cada cliente
* **SALDO LÍMITE**: cantidad de crédito otorgado en dólares
* **SEXO**: Género del cliente (1 = masculino, 2 = femenino)
* **EDUCACIÓN**: (1 = escuela de posgrado, 2 = universidad, 3 = escuela secundaria, 0-4 = otros, 5 = desconocido, 6 = desconocido)
* **MATRIMONIO**: Estado civil (1 = casado, 2 = soltero, 3 = divorciado, 0 = otros)
* **EDAD**: edad en años
* **PAGAR\_0**: Estado de reembolso en septiembre de 2005 (-2 = sin consumo, -1 = pagado, 0 = uso del crédito rotativo, 1 = retraso en el pago durante un mes, 2 = retraso en el pago durante dos meses, ... 8 = retraso en el pago durante ocho meses)
* **PAGAR\_2**: Estado de reembolso en agosto de 2005 (escala igual a PAGAR\_0)
* **PAGAR\_3**: Estado de reembolso en julio de 2005 (escala igual a PAGAR\_0)
* **PAGAR\_4**: Estado de reembolso en junio de 2005 (escala igual a PAGAR\_0)
* **PAGAR\_5**: Estado de reembolso en mayo de 2005 (escala igual a PAGAR\_0)
* **PAGAR\_6**: Estado de reembolso en abril de 2005 (escala igual a PAGAR\_0)
* **CUENTA\_AMT1**: Monto del estado de cuenta en septiembre de 2005 (dólar NT)
* **CUENTA\_AMT2**: Monto del estado de cuenta en agosto de 2005 (dólar NT)
* **CUENTA\_AMT3**: Monto del estado de cuenta en julio de 2005 (dólar NT)
* **CUENTA\_AMT4**: Monto del estado de cuenta en junio de 2005 (dólar NT)
* **CUENTA\_AMT5**: Monto del estado de cuenta en mayo de 2005 (dólar NT)
* **CUENTA\_AMT6**: Monto del estado de cuenta en abril de 2005 (dólar NT)
* **PAGAR\_AMT1**: Monto del pago anterior en septiembre de 2005 (dólar NT)
* **PAGAR\_AMT2**: Monto del pago anterior en agosto de 2005 (dólar NT)
* **PAGAR\_AMT3**: Monto del pago anterior en julio de 2005 (NT dólar)
* **PAGAR\_AMT4**: Monto del pago anterior en junio de 2005 (dólar NT)
* **PAGAR\_AMT5**: Monto del pago anterior en mayo de 2005 (NT dólar)
* **PAGAR\_AMT6**: Monto del pago anterior en abril de 2005 (dólar NT)
* **pago predeterminado el próximo mes**: pago predeterminado (1 = sí, 0 = no)

## Problema

Un incremento en las tasas de incumplimiento de los clientes es malo para Credit One, ya que su negocio está aprobando préstamos para clientes de primera entrada. Es probable que esto resulte en la pérdida de los clientes comerciales de Credit One.

## Preguntas para investigar

1. ¿Cómo asegurar que los clientes pagarán sus préstamos?

* En función de los costos asociados con los valores predeterminados de los clientes, podemos ajustar la precisión y la recuperación. Por ejemplo, ajustando el nivel de umbral para la probabilidad de la predicción predeterminada y la etiqueta final de “No Predeterminada” y “Predeterminada” producida por el modelo.
* Como se describe en el pregunta [2] de este apartado, basado en lecciones aprendidas “No podemos controlar los hábitos de gasto de los clientes”. Para este particular, se recomienda al formalizarse los créditos, inscribir al cliente en el débito automático, ofreciendo algunos beneficios agregados, como por ejemplo, pagar temprano o más dinero sin penalidades de ningún tipo. Lo anterior puede ayudar a garantizar que los pagos se realicen a tiempo. Ofreciendo al cliente hasta reducciones en las tasas de interés, al formalizar bajo este tipo de normativas. Entre mayor el riesgo de no cumplimiento de pago, mayor deben ser las restricciones al iniciar un proceso de crédito.

1. ¿Podemos aprobar clientes con alta certeza?

* Si, considerando los datos de clientes actuales el atributo”pago predeterminado el próximo mes”, predeterminados en “Por Defecto”.
* Considerando las características que según el análisis realizado, representan con mayor certeza el cumplimiento de pago de las deudas adquiridas, como lo fueron: bill\_amt1, pay\_0, limit\_bal, pay\_amt2 y pay\_amt4.

## Conclusiones

Se investigaron los datos, se verificó el desequilibrio de datos, y se visualizaron las características para comprender la relación entre las diferentes características.

Posteriormente investigué cinco modelos predictivos:

Comenzamos con el clasificador Random Forest, para el cual obtuvimos un puntaje de AUC de 0.66.

Para el clasificador Random Forest también intenté con OneHotEncoder (identificar cada característica que se está codificando), reemplazando las características categóricas con valores dummified (introduciendo un dummy variable para cada categoría). La puntuación de AUC no mejoró significativamente en este caso.

Utilicé luego un modelo AdaBoostClassifier (iteraciones), consiguiendo una puntuación de AUC más baja (0,65).

Seguí probando con un modelo CatBoostClassifier (clasificador en la forma de árboles de decisión), obteniendo un puntaje AUC más bajo (0.66).

Luego experimenté con un modelo XGBoost (algoritmo de árboles aumentados de gradientes), para el cual el puntaje AUC obtenido fue de 0.77.

Finalmente probé los datos en un modelo LightGBM (modelo de predicción débil para nuestro problema de clasificación). Usando divisiones de validación de entrenamiento y validación cruzada para evaluar la efectividad del modelo para predecir el valor objetivo, es decir, detectar si un cliente de tarjeta de crédito no pagará el próximo mes. Con ambos métodos para LightGBM los mejores valores obtenidos de AUC para el conjunto de validación fueron de alrededor de 0,78. Siendo este último modelo la mejor opción encontrada.

## Recomendaciones

A partir del análisis inicial, considerando todas las características y valores del dataframe, es evidente que los valores atípicos (outliers): como los clientes que deben más de 500.000 dólares; deberían analizarse por separado, porque realmente impactan los resultados del análisis incluidos como registros normales en el conjunto de datos, porque sobre exponen el resto de características a un incremento no distribuido en el estudio.

A partir de la característica de “pago predeterminado por defecto” considerar los positivos, durante el proceso de sugerencias a los clientes potenciales, adicionando mayor compromiso de pago a los créditos y clientes de mayor riesgo de incumplir con el pago, basados en la densidad y cantidad de límite de crédito.