Ingeniería en ciencia de datos y matematicas
MA2008B.601 - Análisis numérico para la optimización no-lineal (Gpo 601)
Responsable del curso: Javier Edgardo Garrido Guillén

Informe de resultados

Optimizando la Gestión de Clientes: Insights Clave de Datos y Modelos Predictivos

Saturday 3rd May, 2025
Adara Luisa Pulido Sánchez
Cristobal Medina Meza
Guillermo Villegas Morales
Jorge Eduardo Guijarro Márquez
Paulina Martínez López

Contents

1 Resumen ejecutivo

Gestionar adecuadamente una base de clientes diversa es esencial para asegurar la estabilidad financiera de una organización. Uno de los principales retos es identificar, con anticipación, qué clientes podrían presentar retrasos en sus pagos para así actuar de manera preventiva.

Este proyecto tuvo como objetivo desarrollar herramientas analíticas que ayuden a tomar decisiones informadas, utilizando datos reales sobre los clientes. Se analizaron distintos tipos de información, incluyendo datos personales, historial de pagos, canales utilizados y fechas relevantes.

Primero, se dieron los datos por preparar para asegurar su calidad y claridad. Luego, se fueron analizando en busca de patrones relevantes. Uno de los hallazgos importantes fue que solo el 30% de los clientes tenía un comportamiento de pago puntual, mientras que el 70% no lo tenía. También se verificó que la mayoría de los clientes que se retrasan acostumbran regularizar su situación en el segundo mes de morosidad.

Para anticipar estos comportamientos, se usaron varios modelos de predicción, como XGBoost y LightGBM, que demostraron ser muy precisos (99.5% de acierto). Estas herramientas permiten identificar a los clientes con mayor riesgo de no pagar, lo cual abre la puerta a estrategias de cobranza más efectivas y personalizadas.

En resumen, gracias al uso inteligente de los datos y modelos predictivos, es posible mejorar significativamente la gestión de cobros, reduciendo riesgos y fortaleciendo la relación con los clientes.

2 Introducción

2.1 Contexto

En la actualidad, para las empresas que ofrecen productos o servicios financieros a crédito es todo un reto obtener el cumplimiento de los pagos por parte de los consumidores. A medida que las bases de datos crecen y se hacen más heterogéneas, surge la necesidad de aprovechar herramientas innovadoras para analizar mejor los patrones de comportamiento de los clientes, en particular los relacionados con la puntualidad o el impago. La falta de eficacia en la gestión de estos factores puede provocar enormes pérdidas y afectar a la longevidad de las empresas.

2.2 Objetivos

El objetivo principal de la investigación es desarrollar herramientas analíticas que permitan predecir los patrones de pago de los consumidores. Mediante el uso de análisis de tendencias basados en datos históricos, la investigación pretende desarrollar modelos predictores con el fin de determinar aquellos clientes con mayor probabilidad de impago, permitiendo así una toma de decisiones más eficaz y estratégica en las estrategias de cobro.

2.3 Justificación

Este esfuerzo es muy relevante porque permite transformar amplios conjuntos de datos en percepciones significativas para la empresa. Gracias a la capacidad de predecir posibles comportamientos de alto riesgo, es posible elaborar estrategias de cobro más eficaces y adaptadas y garantizar aún más la eficiencia de los procesos y mejorar las relaciones con los clientes. En general, el análisis no solo pretende reducir la morosidad, sino fomentar prácticas más sostenibles y basadas en datos.

3 Descripción de los datos

3.1 Fuente de los datos

Los datos provienen de un archivo de texto con clientes de Bradescard (COLL_TEC_CONSOLIDADO.txt), el cual contiene información histórica de clientes con tarjetas de crédito, prestamos individuales, etc. El dataset incluye registros de pago, comportamiento crediticio, datos demográficos y otros datos que describen el comportamiento de los clientes.

3.2 Estructura

La base de datos contiene 1,289,881 registros con 98 variables, incluyendo:

Variables demográficas: Género, Estado, CP, Fecha de nacimiento Variables crediticias: Saldo total, Saldo mensual, Pago mínimo, Utilización, Ciclo de atraso Variables históricas: Información de saldos, pagos y comportamiento en los últimos 6 meses (M1-M6) Variables de comportamiento: Score de pago, Behavior (indicador de comportamiento crediticio) Variable objetivo: Variable binaria (0,1) que clasifica a los clientes según su historial de pago

3.3 Definición del target y características

La variable objetivo (”Variable_objetivo”) es una variable binaria donde: 0: Representa clientes que no realizan sus pagos a tiempo. 1: Representa clientes que sí realizan sus pagos puntualmente.

3.4 Calidad de los datos

Se detectaron varios problemas en la calidad de los datos:

- Valores faltantes en múltiples columnas, especialmente en variables relacionadas con pagos históricos
- Columnas con tipos de datos mixtos, que requirieron conversión y limpieza
- Presencia de valores atípicos en múltiples variables numéricas.

3.5 Resumen estadístico

- Alta variabilidad en los montos de saldo (promedio de 5,290 con desviación estándar de 5,347)
- Utilización promedio de 8.05 (con valores extremos que llegan hasta 3,970)
- Score_pago promedio de 4.19 (en escala de 0 a 18)
- La mayoría de los clientes tienen un Ciclo_Atraso de 2 (mediana), con un rango de 2 a 4

4 Metodología

4.1 Procesamiento de los Datos

4.1.1 Limpieza

Para tratar con inconsistencias dentro de la base de datos, e realizaron las siguientes operaciones de limpieza:

- Conversión de fechas al formato datetime utilizando formato día-mes-año
- Tratamiento de valores faltantes mediante imputación (con 0, únicamente para variables de utilización y ciclo de atraso)
- Filtrado de clientes inactivos: se eliminaron clientes con promedio de utilización menor al 10
- Eliminación de outliers usando el método IQR para variables de utilización
- Codificación de variables categóricas usando LabelEncoder

4.1.2 Transformaciones

De igual manera, se realizaron los siguientes procesos de transformación:

- Conversión de columnas de tipo string a formato bytes para manejo eficiente
- Creación de nuevas características calculando promedios de utilización por cliente
- Aplicación de escalado estándar (StandardScaler) a las variables numéricas antes del modelado
- Reducción de dimensionalidad mediante PCA para visualización

4.2 Análisis Exploratorio de los Datos

Distribución de la variable objetivo: 68.8% son clientes que no pagan a tiempo (0) y 31.2% son clientes que pagan puntualmente

Se observaron correlaciones significativas entre variables de utilización y ciclo de atraso.

Algunos histogramas de las variables numéricas revelaron distribuciones sesgadas, especialmente en la variable de ”Utilización”.

Realizando análisis de variables categóricas se encontraroon patrones relacionados con el canal de pago y tipo de producto.

4.3 Estudio de Segmentación

- Para el estudio de segmentación se utilizó un enfoque en el que comparamos el desempeño de diferentes algoritmos de clustering, con el objetivo de identificar la técnica que mejor agrupa a los clientes segúnsu comportamiento. Para ello, seguimos una metodología definida:
- Selección de variables relevantes: Se eligieron principalmente variables relacionadas con el comportamiento de pago, utilización y ciclo de atraso que nos daban la información relevante para agrupar a los clientes.
- Preparación de datos para clustering: Se utilizaron métodos como **StandardScaler** para normalización de las variables y métodos de reducción de dimensiones como PCA para visualización de resultados.
- Implementación de modelos de Clasificación como K-Means, Gaussian Mixture Model (GMM) y DBSCAN para realizar la segmentación.

4.4 Feature Selection

Para seleccionar las variables se realizó un análisis exploratorio de datos que incluyó el estudio de correlaciones y visualizaciones de distribución. Se priorizaron aquellas variables con alta correlación con la variable objetivo (*Variable_objetivo*), eliminando aquellas con alta multicolinealidad o elevado porcentaje de valores nulos.

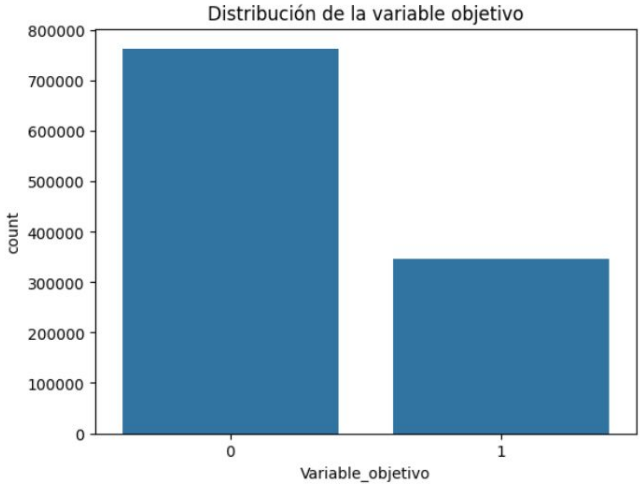

Figure 1: Canal Pago M2.

Se identificó un claro desbalance en los datos: aproximadamente el 70% de los clientes tienen comportamiento moroso. Este desequilibrio se trató durante el modelado, ponderando las clases.

Las variables seleccionadas incluyeron:

- Utilización
- Ciclo de Atraso
- Promedios mensuales (M1–M6)
- Canales de pago y variables temporales

4.5 Modelos

4.5.1 Justificación de la elección del modelo o técnica

Se utilizaron modelos de ensamble como **XGBoost** y **LightGBM** debido a su alta capacidad predictiva, eficiencia computacional y tolerancia a datos faltantes. Además, permiten capturar relaciones no lineales entre variables y realizar interpretación de importancia de características. Para la segmentación de clientes, se emplearon algoritmos no supervisados como **K-Means**, **Gaussian Mixture Model (GMM)** y **DBSCAN**, con el fin de identificar patrones ocultos de comportamiento.

4.5.2 Supuestos e hiperparámetros

- En los modelos supervisados, se ajustaron hiperparámetros clave como la profundidad máxima, número de árboles, tasa de aprendizaje y parámetros de regularización mediante búsqueda en malla (`GridSearchCV`).
- Para K-Means se asumió esfericidad en los grupos; GMM permitió formas elípticas; y DBSCAN, al no requerir número de clusters, se calibró mediante `eps` y `min_samples`.

4.5.3 Implementación del modelo y resultados preliminares

Los modelos predictivos lograron un desempeño sobresaliente, con una precisión del 99.5% en la predicción del comportamiento de pago. La segmentación permitió diferenciar subgrupos de clientes según su nivel de riesgo, lo que abre la posibilidad a estrategias de cobranza diferenciadas y más eficaces.

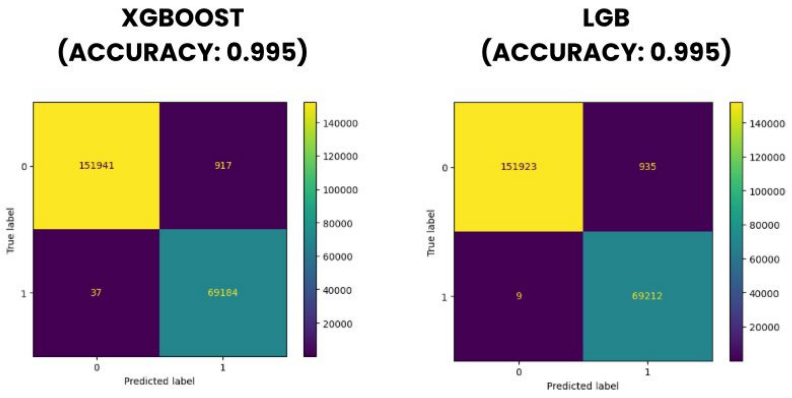

Figure 2: Modelos utilizados.

4.6 Evaluación

4.6.1 Métricas utilizadas

Para evaluar el desempeño de los modelos se utilizaron las siguientes métricas:

- **Precisión (Accuracy)**: ambos modelos alcanzaron un 99.5%, lo que indica un alto porcentaje de predicciones correctas.
- **Recall**: particularmente importante en la clase minoritaria (clientes puntuales), ya que mide la capacidad del modelo para identificar correctamente los verdaderos positivos.
- **F1-Score**: se calculó para equilibrar precisión y recall en un único valor.
- **Curva ROC-AUC**: utilizada para evaluar la capacidad discriminativa del modelo en todos los umbrales posibles. El área bajo la curva fue cercana a 1.

- **Matriz de confusión**: mostró que el modelo cometió muy pocos errores tanto en verdaderos positivos como en verdaderos negativos, lo que indica robustez en ambas clases.

4.6.2 Hallazgos clave

A partir del análisis de desempeño y las gráficas obtenidas, se derivaron los siguientes hallazgos clave:

- Existe un fuerte desbalance de clases en la variable objetivo: solo el 30% de los clientes son puntuales (clase 1), mientras que el 70% presentan impagos (clase 0). Esto se evidenció en la distribución inicial y se compensó en el modelado mediante ponderación de clases.
- **Ambos modelos supervisados (XGBoost y LightGBM)** mostraron un desempeño sobresaliente, con una precisión del 99.5%. Las matrices de confusión revelaron que los errores son mínimos, incluso en la clase minoritaria, donde es más común tener falsos negativos.
- **El modelo LightGBM presentó menor tasa de falsos negativos** (solo 9 clientes cumplidos mal clasificados como morosos), lo que lo hace ligeramente más adecuado para aplicaciones donde es crítico no castigar erróneamente a buenos pagadores.
- Se utilizó la función `predict_proba` para generar probabilidades individuales de impago por cliente. Esta herramienta permite priorizar esfuerzos de cobranza y generar alertas proactivas.
- El análisis de canales de pago por clase reveló diferencias marcadas en el comportamiento según el punto de pago utilizado. Algunos canales como *C&A*, *Bodega* y *GCC* mostraron mayor concentración de clientes morosos en comparación con otros.
- El comportamiento de pago cambia con el tiempo: se detectó que una gran parte de los clientes que se regularizan lo hacen en el mes 2 (M2) tras entrar en morosidad. Esta información puede aprovecharse para diseñar estrategias de intervención temprana.

Las siguientes gráficas apoyan visualmente estos hallazgos y respaldan las decisiones estratégicas propuestas:

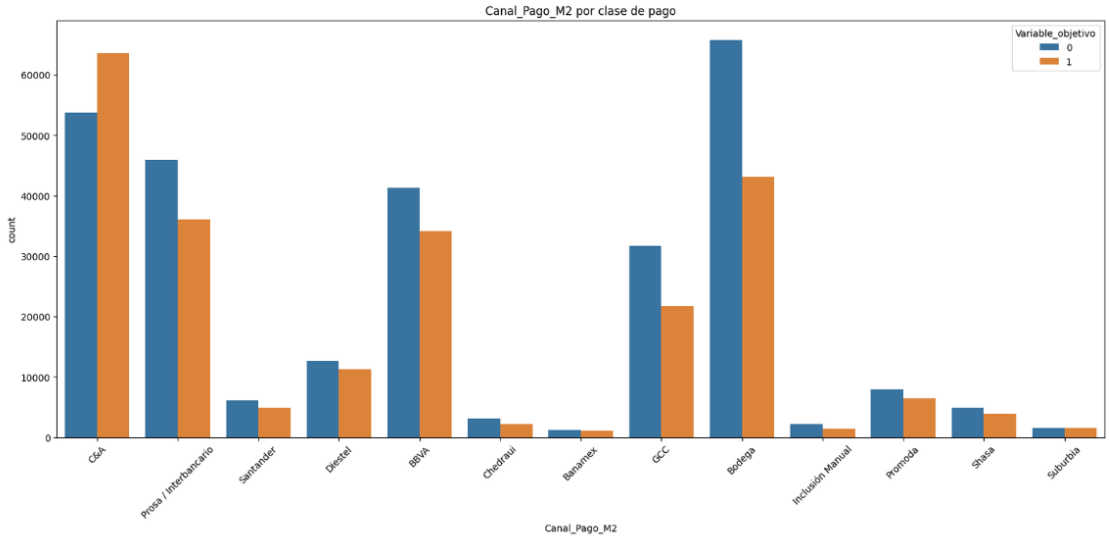

Figure 3: Distribución por canal de pago en mes 2 (M2)

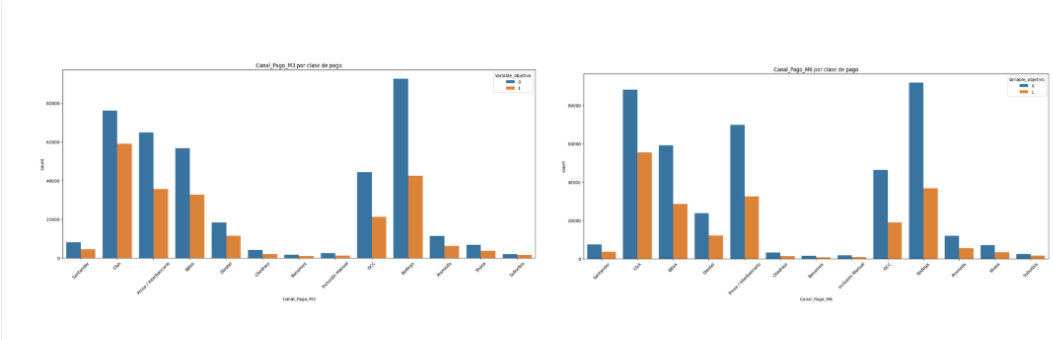

Figure 4: Comparativa de canales de pago en el mes 3 y el mes 6 (M3 y M6) por clase objetivo