CONSTRUCCIÓN DEL DATASET FINAL Y LA DEFINICIÓN DE LA VARIABLE OBJETIVO (Y)

1. Introducción

Esta sección tiene el proposito de describir el proceso seguido para la construcción del dataset final y la definición de la variable objetivo (Y) del modelo predictivo de morosidad estudiantil. Todas las decisiones metodológicas fueron tomadas con acompañamiento de los stakeholders responsables de crédito, cartera, y el Director financiero. El detalle del contenido de cada base de datos utilizada, así como las transformaciones y procesos de limpieza aplicados, puede consultarse en el diccionario de datos correspondiente.

2. Unidad de análisis y estructura del dataset

Cada fila del dataset final representa un crédito otorgado a un estudiante en un periodo académico específico, identificado mediante la clave IdEstudiante_Periodo_Credito. Esta estructura es adecuada porque un mismo estudiante puede presentar características distintas en cada crédito que adquiere. Es posible que en un periodo el crédito sea asumido por un acudiente y en otro por el propio estudiante, o que existan créditos simultáneos en diferentes programas, niveles o modalidades. Los riesgos asociados a cada operación son independientes, por lo que cada crédito debe modelarse como una observación separada. Este enfoque es consistente con las prácticas utilizadas en modelos de riesgo financiero, donde la unidad de análisis es la operación específica y no el individuo.

3.Fuentes de información integradas

Para la conformación del dataset se integraron cuatro bases de información principales: créditos, cartera, scoring externo y datos académicos.

- La base de créditos aportó los elementos estructurales del crédito otorgado a cada estudiante.
- La base de cartera permitió identificar el estado de las cuotas asociadas al crédito, clasificando aquellas que se encuentran vencidas y las que están por vencer.
- La base de scoring externo proporcionó el puntaje de riesgo inicial asignado por un tercero antes de la aprobación del crédito.
- La base académica permitió complementar la información con variables relativas al comportamiento y contexto del estudiante dentro de la institución.

Para más detalles sobre la estructura, campos y depuración de cada una de estas fuentes, se remite al diccionario de datos.

4. Lógica acordada con stakeholders para la definición de la variable objetivo

En las sesiones de trabajo con los stakeholders se acordó que el objetivo del modelo es clasificar los créditos en niveles de riesgo: bajo, medio y alto. Para lograrlo, la variable objetivo debía reflejar el comportamiento real del crédito hasta la fecha de corte sin incorporar información futura que pudiera generar fuga de información. El análisis permitió concluir que la fuente adecuada para este cálculo era la base de cartera, dado que allí se registra el estado de las cuotas que ya debían haberse pagado.

5. Metodología para el cálculo de la variable objetivo (Y)

Para definir la Y se acordó utilizar únicamente las cuotas vencidas y las cuotas por vencer. Las cuales no forman parte del modelo y solo se emplean temporalmente para definir nuestra variable respuesta. El cálculo se realiza mediante la siguiente expresión:

$$Y = \frac{\text{Cuotas vencidas}}{\text{Total de cuotas pactadas} - \text{Cuotas por vencer}}$$

Este planteamiento permite excluir del cálculo todas las cuotas futuras que aún no han alcanzado su fecha de vencimiento, evitando distorsiones en la medición del comportamiento de pago y las cuotas actuales sin pagar debido a que esta es literalmente la respuesta que el modelo intenta predecir. Una vez calculada la Y, toda la información relacionada con la cartera se elimina del dataset final con el fin de impedir el ingreso de variables que reflejen comportamientos posteriores al origen del crédito.

6. Eliminación de variables susceptibles de fuga de información

Para garantizar que el modelo sea estrictamente predictivo, se excluyeron del dataset final todas las variables derivadas del comportamiento de pago del crédito actual, incluyendo cuotas vencidas, cuotas por vencer, estados de pago y fechas asociadas a la cartera. Estas variables fueron utilizadas exclusivamente para el cálculo inicial de la Y y no se incorporan en el set de entrenamiento. Con ello se evita la fuga de información y se asegura que el modelo solo utilice características conocidas en el momento de aprobación del crédito.

7. Conformación del dataset final

Con base en las decisiones anteriores, el dataset final contiene únicamente variables permitidas para predicción: características del crédito al origen, información académica del estudiante, datos demográficos y el scoring externo. Las variables relacionadas con la cartera fueron excluidas tras el cálculo de la Y. El detalle de los campos conservados y eliminados puede consultarse en el diccionario de datos, donde se documentan las reglas aplicadas para cada variable y la justificación de su inclusión o exclusión.

8. Conclusión

El dataset final y la definición de la variable objetivo fueron construidos de manera rigurosa, transparente y alineada con los requerimientos del negocio. Se contó con la participación activa de los stakeholders para asegurar que las decisiones metodológicas permitieran un modelo válido, ético y útil para la clasificación de riesgo de morosidad. La elección de la unidad de análisis, la integración de múltiples fuentes de información y la eliminación de variables incompatibles con un modelo predictivo permiten asegurar un conjunto de datos consistente y listo para la construcción del modelo de morosidad. El enfoque adoptado garantiza que las predicciones serán éticamente correctas, técnicamente válidas y coherentes con las necesidades institucionales.