Sebastián Ruiz

30 de Marzo del 2025

Preparación de Datos: Proyecto de Optimización del Proceso de Evaluación de Solicitantes

1. Entendimiento del Negocio

El objetivo principal de este proyecto es mejorar el proceso de evaluación de solicitantes de préstamos con el fin de minimizar la cantidad de préstamos incumplidos. Esto permitirá a la institución financiera reducir pérdidas, optimizar la asignación de recursos y mejorar la rentabilidad general. Para lograrlo, se analizarán los datos históricos de préstamos y se desarrollará un modelo predictivo que identifique a los solicitantes con mayor probabilidad de incumplimiento.

Preguntas clave:

- ¿Qué características de los solicitantes están más asociadas con el incumplimiento de préstamos?
- ¿Cómo se puede predecir el riesgo de incumplimiento antes de otorgar un préstamo?
- ¿Qué ajustes se pueden realizar en el proceso de evaluación para reducir el riesgo?

2. Entendimiento de los Datos

Los datos fueron descargados de Kaggle en la siguiente dirección: https://www.kaggle.com/datasets/joebeachcapital/loan-default/

El dataset utilizado contiene información histórica sobre préstamos, incluyendo características de los solicitantes, detalles del préstamo y su estado final (pagado, incumplido, etc.). Las columnas más relevantes incluyen:

loan_amnt: Monto solicitado por el prestatario.

int_rate: Tasa de interés del préstamo.

annual_inc: Ingreso anual del prestatario.

emp_length: Duración del empleo del prestatario.

loan_status: Estado final del préstamo (pagado, incumplido, etc.).

dti: Relación deuda-ingreso del prestatario.

Se identificaron valores faltantes, outliers y columnas irrelevantes que serán tratados en la etapa de preparación de datos.

---

3. Preparación de Datos

La preparación de datos es una etapa crítica en este proyecto, ya que garantiza que los datos estén limpios, relevantes y listos para el análisis y modelado. A continuación, se describen los pasos realizados:

---

1. Eliminación de Columnas Irrelevantes

Se eliminaron columnas que no aportan valor directo al análisis o que son redundantes:

- **zip_code**, **id**, **member_id**: Identificadores únicos o información demasiado específica que no contribuyen al análisis.
- **Columnas relacionadas con el historial crediticio**:
    - **earliest_cr_line**: Fecha de la primera línea de crédito, no relevante para el objetivo del proyecto.
    - **inq_last_6mths**: Número de consultas de crédito en los últimos 6 meses, redundante con otras métricas.
    - **revol_bal** y **revol_util**: Información sobre el uso de crédito rotativo, ya resumida en otras variables como dti.
    - **total_acc**: Número total de cuentas de crédito, no directamente relevante.
- **term**: Reemplazada por una nueva columna binaria long_term que indica si el préstamo es a largo plazo (60 meses).

---

2. Manejo de Valores Faltantes

Se lidió así con los valores faltantes en las siguientes columnas:

- **emp_length**: Los valores faltantes se reemplazaron con 0, asumiendo que representan personas desempleadas.
- **mths_since_last_delinq**: Los valores faltantes se reemplazaron con 240, indicando que no hubo morosidad previa.
- **last_pymnt_d**: Los valores faltantes se reemplazaron con issue_d para calcular métricas derivadas como time_to_delinq.

---

3. Creación de Nuevas Columnas

Se generaron nuevas columnas para enriquecer el análisis:

- **time_to_delinq**: Calcula el tiempo (en días) entre la emisión del préstamo (issue_d) y el último pago (last_pymnt_d) para préstamos que no están completamente pagados.
- **long_term**: Columna binaria que indica si el préstamo tiene una duración de 60 meses (1) o 36 meses (0).

- **employment_verified**: Columna binaria que convierte verification_status en 0 para Not Verified y 1 para cualquier otro valor (Verified o Source Verified).

---

4. Eliminación de Outliers

Se identificaron y eliminaron valores extremos en la columna annual_inc (ingreso anual). Los ingresos se limitaron al percentil 99 para evitar que valores atípicos distorsionen el análisis.

---

5. Consolidación de la Variable Objetivo

La columna loan_status se consolidó en valores Fully Paid, Charged Off, Late y Current.

Además, se eliminó la columna repay_fail, ya que su información está contenida en loan_status.