Proyecto Inteligencia Artificial, Informe #2

Daniel Lujan Agudelo

Emanuel López Higuita

Santiago Rivera Montoya

Modelos y Simulación de Sistemas I

Facultad de Ingeniería, Universidad de Antioquia

Octubre 2023

1. Introducción

Home Credit es una institución financiera internacional que se enfoca en el préstamo de dinero a personas con poco o nulo historial crediticio.

Con el fin de buscar mayor rentabilidad, se requiere un modelo de Machine Learning (ML) que prediga con qué probabilidad un cliente solicitante de un crédito, dada una serie de datos personales y financieros, pagará su deuda debidamente.

2. Exploración de datos

La empresa dispuso a través de una competición de Kaggle, un conjunto de datos de sus clientes para entrenar un modelo de ML.

El dataset cuenta con un total de 307,511 registros de clientes, para los que se tienen 122 variables o columnas.

El dataset cuenta con un total de 122 variables, de las cuales 106 son continuas y 16 son categóricas. Estas últimas se describen a continuación (Tabla 1):

NAME_CONTRACT_TYPE	Identificación si el préstamo es en efectivo o revolvente.
CODE_GENDER	Género del cliente.
FLAG_OWN_CAR	Marca si el cliente posee un carro.
FLAG_OWN_REALTY	Marca si el cliente posee una casa o un piso.
NAME_TYPE_SUITE	Quien acompañaba al cliente cuando solicitaba el préstamo.
NAME_INCOME_TYPE	Tipo de ingreso del cliente (empresario, trabajador, baja por maternidad, etc).
NAME_EDUCATION_TYPE	Nivel de educación más alto que alcanzó el cliente.
NAME_FAMILY_STATUS	Estado familiar del cliente.

NAME_HOUSING_TYPE	Cuál es la situación de vivienda del cliente (de alquiler, viviendo con los padres, etc).
OCCUPATION_TYPE	Qué tipo de ocupación tiene el cliente.
FONDKAPREMONT_MODE	Información normalizada sobre el edificio donde vive el cliente.
WALLSMATERIAL_MODE	Información normalizada sobre el edificio donde vive el cliente.
EMERGENCYSTATE_MODE	Información normalizada sobre el edificio donde vive el cliente.
WEEKDAY_APPR_PROCESS_START	En qué día de la semana el cliente solicitó el préstamo.
ORGANIZATION_TYPE	Tipo de organización en la que trabaja el cliente.

Tabla 1

Además, es necesario analizar la información faltante en las columnas. Algunos datos a destacar:

- La columna OWN_CAR_AGE, que proporciona la edad del vehículo de la persona **en caso de tenerlo**, tiene sólo 34% de los datos.
- La columna OCCUPATION_TYPE, que proporciona la ocupación de la persona, tiene un 69% de los datos. Nótese que la información que proporciona esta columna es fundamental para la predicción.

3. Preprocesamiento de datos

Antes de usar algún modelo, es necesario normalizar/procesar los datos, eliminando información irrelevante, y reemplazando la información que falte de la forma más adecuada posible.

3.1. Rellenado de datos faltantes

El primer proceso por el que pasan los datos es por el rellenado de los datos faltantes.

Para el caso de las variables continuas, se decidió que todos los datos faltantes serían reemplazados con el valor medio de la columna correspondiente.

Mientras que para las variables categóricas, para esta primera iteración, se reemplazaron los datos faltantes con la categoría más común en la columna. Esto puede no ser adecuado para algunas variables.

Por ejemplo, en la columna OCCUPATION_TYPE, el valor más común es el de Empleado (Ver Figura 1), sin embargo, lo ideal sería agrupar a las personas con el dato faltante en una nueva categoría 'Desempleado'.

	OCCUPATION_TYPE
Laborers	151577
Sales staff	32102
Core staff	27570
Managers	21371
Drivers	18603
High skill tech staff	11380
Accountants	9813
Medicine staff	8537
Security staff	6721
Cooking staff	5946
Cleaning staff	4653
Private service staff	2652
Low-skill Laborers	2093
Waiters/barmen staff	1348
Secretaries	1305
Realty agents	751
HR staff	563
IT staff	526

(Figura 1)

3.2. Variables categóricas a numéricas

Para transformar las variables categóricas a números y poder usarlas como datos de entrenamiento para el modelo, se usó One-Hot Encoding sobre todas las variables tabuladas en la Tabla 1.

3.3. Posibles variables innecesarias

Dentro del dataset, existen algunas variables que pueden no ser de utilidad para predecir si un cliente pagará o no el crédito, y que, en cambio, inducen error en el modelo. Una vez identificadas estas columnas, es conveniente eliminarlas del conjunto de datos de entrenamiento.

Para la primera predicción, no se ha quitado ninguna columna. Sin embargo, se han identificado algunas variables que posiblemente deban ser eliminadas:

- En qué día de la semana el cliente solicitó el préstamo (WEEKDAY_APPR_PROCESS_START).
- Quien acompañaba al cliente cuando solicitaba el préstamo (NAME_TYPE_SUITE).

4. Primer modelo predictivo

La primera predicción hecha se realizó usando el modelo **Random Forest Classifier** usando 100 árboles (o estimadores).

4.1. Resultados

Usando el dataset procesado y el dataset para pruebas que la competición incluye, se obtuvieron los siguientes resultados:

SK_ID_CURR	100001	100005	100013	100028	100038	100042	100057	100065	100066	100067	...
TARGET	0.11	0.11	0.08	0.06	0.1	0.05	0.02	0.05	0.06	0.23	...

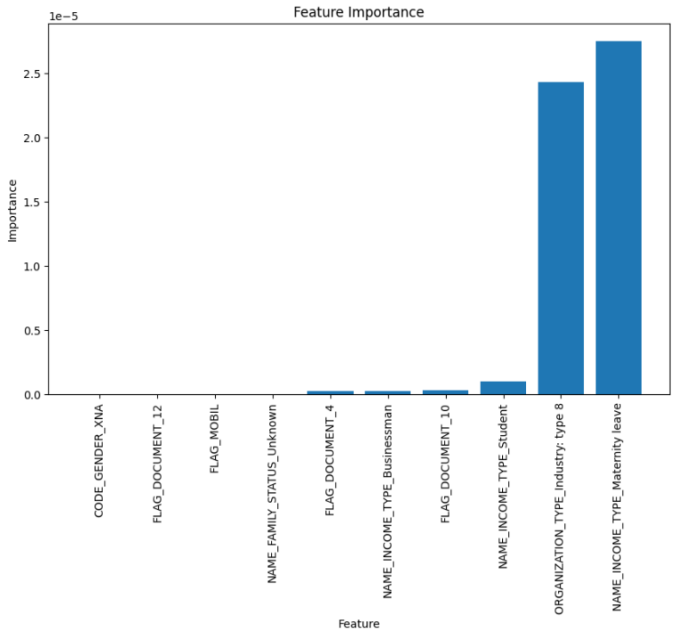
(Figura 2)

Dónde TARGET es la variable objetivo y significa la probabilidad de que el cliente pague su deuda.

El puntaje obtenido usando el **área bajo la curva ROC** como métrica de desempeño, al subir estos resultados a Kaggle fue de ≈ 69.6%

Según el modelo entrenado, los dos factores más influyentes en la probabilidad de que el cliente pague su deuda son los siguientes (ver Figura 3):

- Si los ingresos de la persona provienen del permiso de maternidad.
- Si el tipo de organización en la que trabaja el cliente es una empresa.



(Figura 3)

Otro dato a destacar sobre los resultados, es que el promedio de la probabilidad de que el cliente pague es, aproximadamente, 8.53%. Este porcentaje es similar al de personas que pagaron su deuda satisfactoriamente en el dataset de entrenamiento (promedio de la variable objetivo): 8.07%.