PRIMERA ENTREGA DE PROYECTO IA

Presentado por:

Sebastian Castro Bolaños

Karol Melissa Reyes Anaya

Jorge Andrés Cardeño Devia

Profesor

Raúl Ramos Pollan

Universidad de Antioquia

Facultad de Ingeniería

Medellín 2023-2

RIESGO DE INCUMPLIMIENTO DE CRÉDITO HIPOTECARIO

1. Descripción del problema

Nuestro propósito en este proyecto es predecir la capacidad de pago de una población no bancarizada a partir del uso de maching learning. En este proyecto se hará uso de datos históricos de solicitudes de préstamos para predecir si un solicitante podrá pagar un préstamo o no. Home Credit busca mejorar su capacidad de evaluar el riesgo de impago y tomar decisiones informadas sobre la aprobación o denegación de solicitudes de préstamo. Esta es una tarea de clasificación supervisada estándar:

Supervisado: Las etiquetas se incluyen en los datos de entrenamiento y el objetivo es entrenar un modelo para que aprenda a predecir las etiquetas a partir de las características.

Clasificación: La etiqueta es una variable binaria, 0 (pagará el préstamo a tiempo), 1 (tendrá dificultades para pagar el préstamo).

2. DataSet Utilizado

Vamos a utilizar el Dataset de kaggle *Home Credit Default Risk (Riesgo de incumplimiento de crédito hipotecario | Kaggle),* que contiene los siguientes datos:

	Dataframe	Filas	Columnas
0	application_train	307511	122
1	POS_CASH_balance	10001358	8
2	bureau_balance	27299925	3
3	previous_application	1670214	37
4	installments_payments	13605401	8
5	credit_card_balance	3840312	23
6	bureau	1716428	17
7	application_test	48744	121

Figura 1. Tamaño de datos proporcionado por la competencia.

En el proyecto, nos limitaremos a utilizar solo los datos contenidos en application_train, pues contienen la variable TARGET, que es la que nos piden predecir en la competencia. Esto nos permitirá establecer una base que luego podremos mejorar.

Las columnas numéricas contenidas en estos datos son las siguientes:

```
datos_ordenados = [
    'AMT_ANNUITY', 'AMT_CREDIT', 'AMT_GOODS_PRICE', 'AMT_INCOME_TOTAL', 'AMT_REQ_CREDIT_BUREAU_DAY',
    'AMT_REQ_CREDIT_BUREAU_HOUR', 'AMT_REQ_CREDIT_BUREAU_MON', 'AMT_REQ_CREDIT_BUREAU_QRT',
    'AMT_REQ_CREDIT_BUREAU_WEEK', 'AMT_REQ_CREDIT_BUREAU_YEAR', 'APARTMENTS_AVG', 'APARTMENTS_MEDI',
    'APARTMENTS_MODE', 'BASEMENTAREA_AVG', 'BASEMENTAREA_MEDI', 'BASEMENTAREA_MODE', 'CNT_CHILDREN',
    'CNT_FAM_MEMBERS', 'COMMONAREA_AVG', 'COMMONAREA_MEDI', 'COMMONAREA_MODE', 'DAYS_BIRTH',
    'DAYS_EMPLOYED', 'DAYS_ID_PUBLISH', 'DAYS_LAST_PHONE_CHANGE', 'DAYS_REGISTRATION',
    'DEF_30_CNT_SOCIAL_CIRCLE', 'DEF_60_CNT_SOCIAL_CIRCLE', 'ELEVATORS_AVG', 'ELEVATORS_MEDI',
    'ELEVATORS_MODE', 'ENTRANCES_AVG', 'ENTRANCES_MEDI', 'ENTRANCES_MODE', 'EXT_SOURCE_1',
    'EXT_SOURCE_2', 'EXT_SOURCE_3', 'FLAG_CONT_MOBILE', 'FLAG_DOCUMENT_10', 'FLAG_DOCUMENT_11',
    'FLAG_DOCUMENT_12', 'FLAG_DOCUMENT_13', 'FLAG_DOCUMENT_14', 'FLAG_DOCUMENT_15', 'FLAG_DOCUMENT_16',
    'FLAG_DOCUMENT_17', 'FLAG_DOCUMENT_18', 'FLAG_DOCUMENT_19', 'FLAG_DOCUMENT_2', 'FLAG_DOCUMENT_20',
    'FLAG_DOCUMENT_21', 'FLAG_DOCUMENT_3', 'FLAG_DOCUMENT_4', 'FLAG_DOCUMENT_5', 'FLAG_DOCUMENT_6',
    'FLAG_DOCUMENT_7', 'FLAG_DOCUMENT_8', 'FLAG_DOCUMENT_9', 'FLAG_EMAIL', 'FLAG_EMP_PHONE', 'FLAG_MOBIL',
    'FLAG_PHONE', 'FLAG_WORK_PHONE', 'FLOORSMAX_AVG', 'FLOORSMAX_MEDI', 'FLOORSMAX_MODE', 'FLOORSMIN_AVG',
    'FLOORSMIN_MEDI', 'FLOORSMIN_MODE', 'HOUR_APPR_PROCESS_START', 'LANDAREA_AVG', 'LANDAREA_MEDI',
    'LANDAREA_MODE', 'LIVE_CITY_NOT_WORK_CITY', 'LIVE_REGION_NOT_WORK_REGION', 'LIVINGAPARTMENTS_AVG',
    'LIVINGAPARTMENTS_MEDI', 'LIVINGAPARTMENTS_MODE', 'LIVINGAREA_AVG', 'LIVINGAREA_MEDI', 'LIVINGAREA_MODE',
    'NONLIVINGAPARTMENTS_AVG', 'NONLIVINGAPARTMENTS_MEDI', 'NONLIVINGAPARTMENTS_MODE', 'NONLIVINGAREA_AVG',
    'NONLIVINGAREA_MEDI', 'NONLIVINGAREA_MODE', 'OBS_30_CNT_SOCIAL_CIRCLE', 'OBS_60_CNT_SOCIAL_CIRCLE',
    'OWN_CAR_AGE', 'REGION_POPULATION_RELATIVE', 'REGION_RATING_CLIENT', 'REGION_RATING_CLIENT_W_CITY',
    'REG_CITY_NOT_LIVE_CITY', 'REG_CITY_NOT_WORK_CITY', 'REG_REGION_NOT_LIVE_REGION',
    'REG_REGION_NOT_WORK_REGION', 'SK_ID_CURR', 'TARGET', 'TOTALAREA_MODE', 'YEARS_BEGINEXPLUATATION_AVG',
    'YEARS_BEGINEXPLUATATION_MEDI', 'YEARS_BEGINEXPLUATATION_MODE', 'YEARS_BUILD_AVG', 'YEARS_BUILD_MEDI',
    'YEARS_BUILD_MODE'
]
```

Figura 2. Columnas numéricas datos application_train.

Y las columnas categóricas son las siguientes:

```
['NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR',
'FLAG_OWN_REALTY', 'NAME_TYPE_SUITE', 'NAME_INCOME_TYPE',
'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE',
'OCCUPATION_TYPE', 'WEEKDAY_APPR_PROCESS_START', 'ORGANIZATION_TYPE',
'FONDKAPREMONT_MODE', 'HOUSETYPE_MODE', 'WALLSMATERIAL_MODE',
'EMERGENCYSTATE_MODE']
```

Figura 3. Columnas categóricas datos application_train.

3. Métrica de desempeño

Como métrica se utiliza la recomendada por la competencia, ROC AUC. Se trata de una métrica de clasificación común conocida como área característica operativa del receptor bajo la curva. La curva de característica operativa del receptor (ROC) representa la tasa de verdaderos positivos versus la tasa de falsos positivos.

Esta métrica está entre 0 y 1 y un mejor modelo obtiene una puntuación más alta, que es lo que se espera obtener.

4. Desempeño deseado

En un principio la idea es entender los datos, y la métrica por la cual se juzgará nuestro modelo, esto con la finalidad de crear un modelo sólido que nos permita mantener una baja cantidad de falsos positivos y falsos negativos como lo mencionamos anteriormente, evitando así, que se puedan ver afectados los resultados en la toma de decisiones de préstamo, siendo esto lo que define la capacidad de evaluar el riesgo de impago y el éxito en la selección de clientes, por eso es importante lograr parametrizar las características bajo las cuales se busca entrenar el modelo, ya que de estas dependerá la predicción y la puntuación, la cual se busca sea alta.