# SEGUNDA ENTREGA DEL PROYECTO

Presentado por:

Aura María Molina Amaya
Valeria González González
Maryely Isabel Rubio de la Cruz

Presentado a:
Raúl Ramos Pollán


**UNIVERSIDAD DE ANTIOQUIA**
**DEPARTAMENTO DE INGENIERÍA INDUSTRIAL**
**INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL PARA LAS CIENCIAS E INGENIERÍAS**
**2023**

## 1. Planteamiento del problema

Actualmente, muchas las personas que desean adquirir nuevos préstamos no cuentan con un historial crediticio suficiente o simplemente nunca han adquirido un crédito en su vida; por lo cual deben acudir a otros medios para obtener el dinero y es allí donde prestamistas que no son acreditados para esta labor sacan provecho a beneficio propio. Por ello, y teniendo en cuenta una variedad de datos alternativos, incluida la información de telecomunicaciones y transaccional, se predecirá cuán capaz es cada solicitante de pagar un préstamo.

### 1.1 Dataset

Se utilizará el dataset de una competencia de Kaggle llamada **Home Credit Default Risk** (https://www.kaggle.com/competitions/home-credit-default-risk ).

Para realizar el modelo, se utilizará el archivo con los datos de entrenamiento (application_train.csv), el cual cuenta con una muestra de 307.511 clientes.

Como lo que se busca predecir es el comportamiento de pago futuro de los clientes a partir de datos de comportamiento crediticio de aplicación, demográficos e históricos; en el dataset, se encuentran columnas con información del cliente como el tipo de préstamo, monto del préstamo, género, edad, si tiene carro y/o casa, número de hijos, nivel educativo, ocupación, ingresos, tipo de ingresos, entre otras, como el cumplimiento de la documentación requerida, etc.

### 1.2 Métricas

Como métrica se utilizará la asignada por la competencia: el área bajo la curva ROC entre la probabilidad prevista y el objetivo observado. El AUROC se calcula como el área bajo la curva ROC. Una curva ROC muestra la compensación entre la tasa positiva verdadera (TPR) y la tasa positiva falsa (FPR) en diferentes umbrales de decisión.

Una curva ROC siempre comienza en la esquina inferior izquierda, es decir, el punto (FPR = 0, TPR = 0) que corresponde a un umbral de decisión de 1 (donde cada ejemplo se clasifica como negativo, porque todas las probabilidades pronosticadas son inferiores a 1)

Una curva ROC siempre termina en la esquina superior derecha, es decir, el punto (FPR = 1, TPR = 1) que corresponde a un umbral de decisión de 0 (donde cada ejemplo se clasifica como positivo, porque todas las probabilidades pronosticadas son mayores que 0)

### 1.3 Variable Objetivo

Como variable objetivo, se tiene TARGET (1 - Cliente con dificultades de pago: tuvo un pago atrasado más de x días en al menos una de las primeras cuotas del préstamo en la muestra, 0 - todos los demás casos).

## 2. Exploración de las variables

### 2.1. Análisis de la variable objetivo

Como se mencionó anteriormente, nuestra variable objetivo, TARGET es una variable booleana, es decir que sólo puede tener dos tipos de entrada respecto al incumplimiento de pago del cliente, 0 (No) o 1 (sí). En la figura 1, se muestra el comportamiento de esta variable. Se aprecia que el 91.93% de los datos corresponden a clientes que sí cumplieron con el pago de un préstamo.

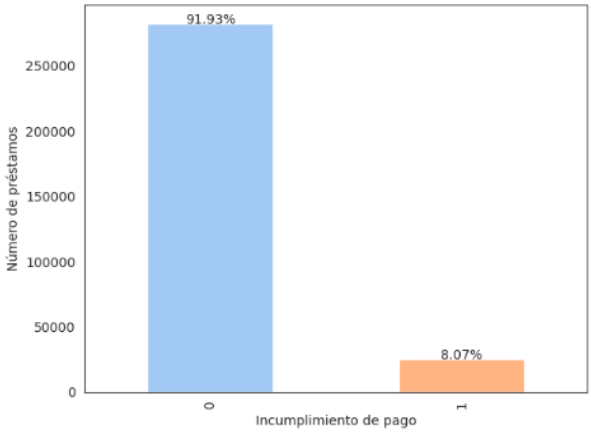


**Figura 1.** *Comportamiento de la variable objetivo.*

### 2.2. Datos faltantes

Una vez realizado el análisis de la variable objetivo, se procedió a revisar cuáles variables tienen mayor porcentaje de datos faltantes, lo cual se muestra en la tabla 1. Podemos ver que las variables con mayor número datos faltantes son *COMMONAREA_MEDI*, *COMMONAREA_AVG* y *COMMONAREA_MODE* con un 69.87% en términos porcentuales.

**Tabla 1.** *Variables con la mayor cantidad de datos faltantes.*

| | Total | Percent |
|---|---|---|
| COMMONAREA_MEDI | 214865 | 69.872297 |
| COMMONAREA_AVG | 214865 | 69.872297 |
| COMMONAREA_MODE | 214865 | 69.872297 |
| NONLIVINGAPARTMENTS_MODE | 213514 | 69.432963 |
| NONLIVINGAPARTMENTS_AVG | 213514 | 69.432963 |

### 2.3. Correlación de variables

En la tabla 2, se muestra la correlación existente entre las diferentes variables con la variable objetivo. Se puede observar que las variables *DAYS_BIRTH* y *REGION_RATING_CLIENT*, son las que mayor correlación tienen con la variable objetivo. Es de aclarar que se redujo la cantidad de variables analizadas a las 7 con mayor correlación, esto debido a la gran cantidad de las éstas que tiene la base de datos.

**Tabla 2.** *Correlación de siete variables con la variable objetivo.*

| | TARGET |
|---|---|
| TARGET | 1.000000 |
| DAYS_BIRTH | 0.078239 |
| REGION_RATING_CLIENT_W_CITY | 0.060893 |
| REGION_RATING_CLIENT | 0.058899 |
| DAYS_LAST_PHONE_CHANGE | 0.055218 |
| DAYS_ID_PUBLISH | 0.051457 |
| REG_CITY_NOT_WORK_CITY | 0.050994 |

### 2.47. Distribución de las variables numéricas

En la figura 4 se muestran las distribuciones de cada variable, aquí podemos observar que son muy pocas las variables que tienen una distribución similar a la distribución normal. También se puede observar que muchas de las variables cuentan con una distribución simétrica hacía la izquierda.

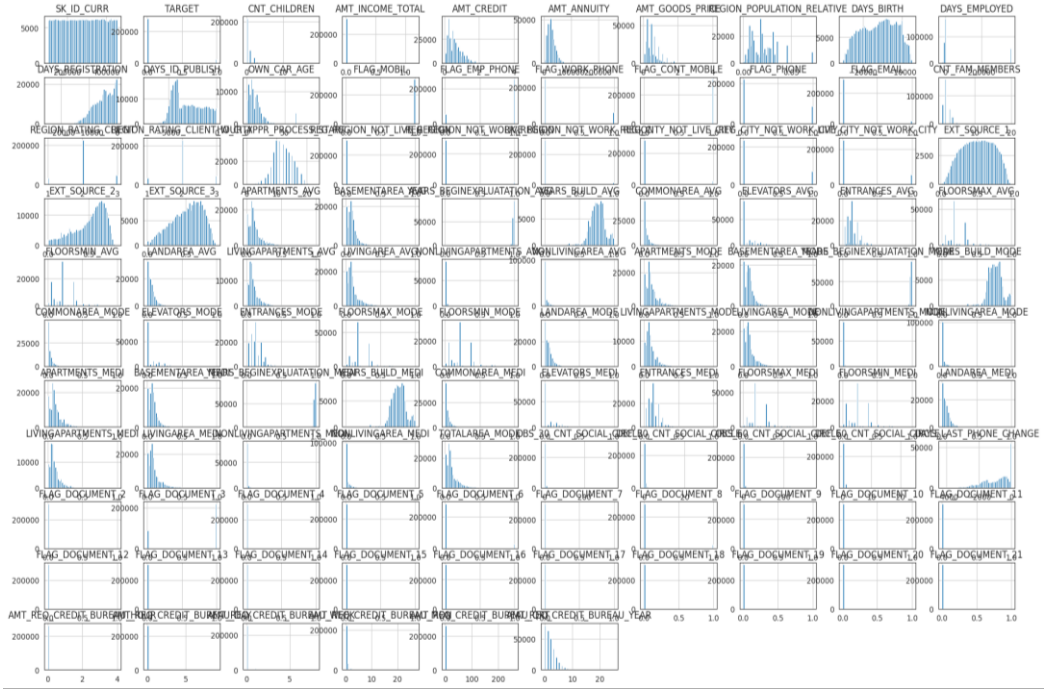


**Figura 4.** *Distribución de las variables.*

En la figura 5, se redujo la cantidad de gráficas de distribución a las siete seleccionadas, las cuales se escogieron las que sus columnas tuviesen mayor correlación con la variable objetivo (mayor a 0.05). Aquí podemos observar que las variables con mayor correlación no siguen una distribución específica.

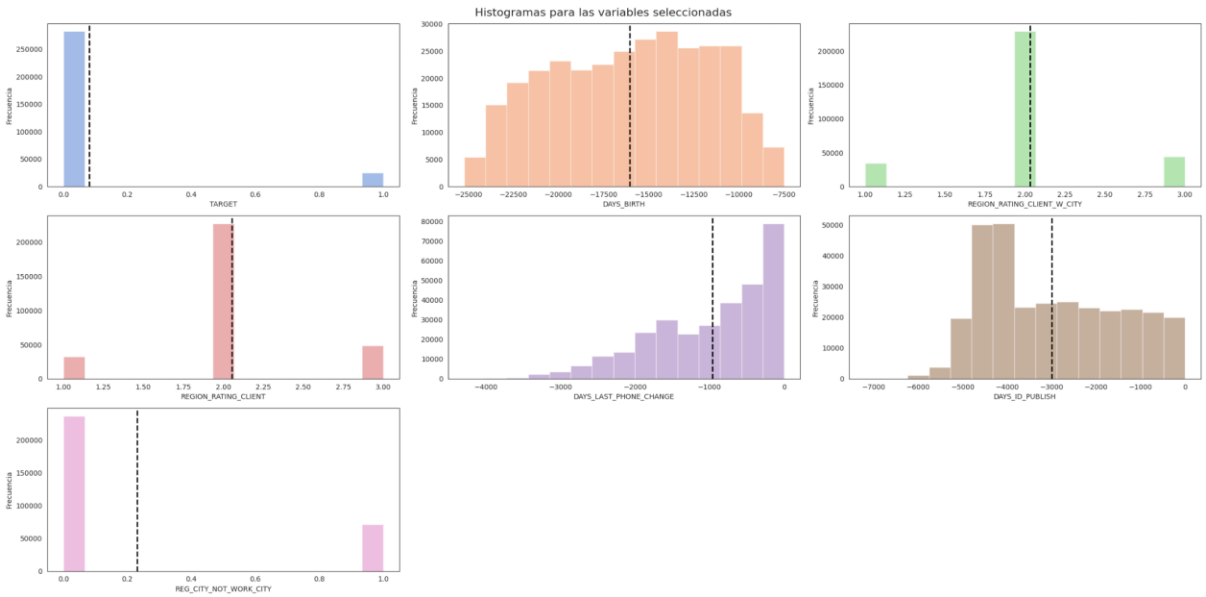


**Figura 5.** *Distribución de las siete variables mayor correlacionadas con la variable objetivo.*

## 3. Tratamiento de datos

- **Eliminación de las columnas con varios datos faltantes:** se procede a eliminar las columnas que cuentan con más del 50% de datos faltantes, ya que se considera que es mucha información la que no se tiene para poder aportar suficiente información al modelo.
- **Relleno de datos faltantes:** se hace un tratamiento del resto de los datos faltantes mediante la imputación por media para las columnas numéricas y de imputación por moda para las columnas categóricas.
- **Transformación de variables categóricas a numéricas:** se hace uso de la Codificación de etiquetas (Label Encoding) y la Codificación One-Hot (One-Hot Encoding) según el número de

valores únicos que tenía cada variable categórica, **con** el fin de que que los modelos de aprendizaje automático puedan trabajar con ellas.

## 4. Modelos

En este caso, se hace uso de modelos tales como Logistic Regression, Random Forest y Gradient Boosting puesto, los cuales son comúnmente utilizados para abordar problemas de clasificación binaria. Además, son modelos rápidos de entrenar y pueden producir un alto rendimiento en la mayoría de los conjuntos de datos. En la figura 6, se muestra el desempeño bajo el criterio de la curva ROC y el área bajo la curva AUC para los tres modelos de clasificación seleccionados. Se observa que Gradient Boosting cuenta con un AUC significativamente mayor a los demás, con un valor correspondiente a 0.75, indicando un rendimiento mayor.

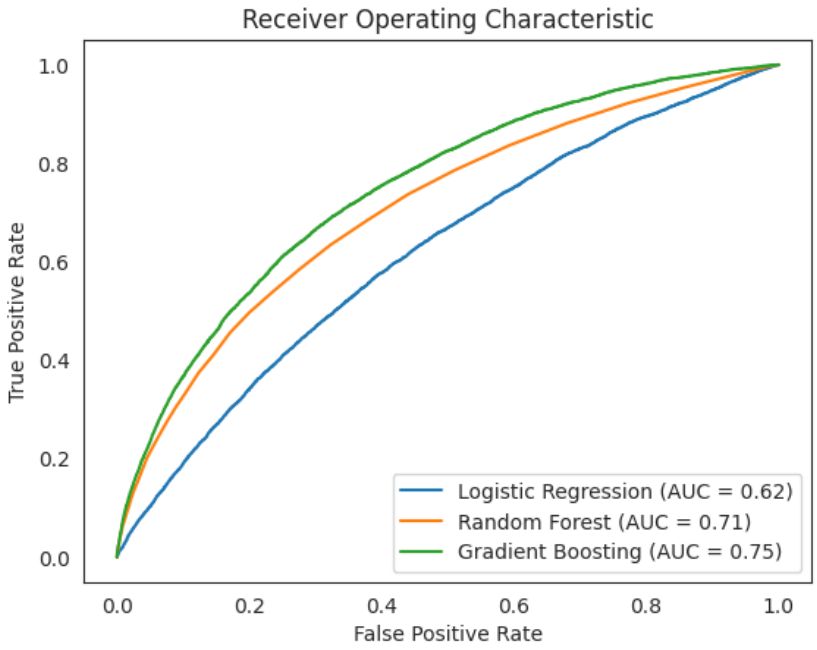


**Figura 6.** *Desempeño de modelos bajo el criterio de la curva ROC.*

## 5. Referencias


Draelos, R (s.f). Measuring Performance: AUC (AUROC) https://glassboxmedicine.com/2019/02/23/measuring-performance-auc-auroc/

Kaggle. (s.f.). Home Credit Default Risk. https://www.kaggle.com/competitions/home-credit-default-risk/overview

Moyete. (2022, 8 marzo). *Curso scikit-learn | Curva ROC | Machine Learning Python 05* [Vídeo]. YouTube. https://www.youtube.com/watch?v=RZiWJIyaQbg