Modelos predictivos para la estimación del riesgo de impago en datos financieros

PABLO BAUTISTA

Contenido

EXPLORACIÓN DE LOS DATOS:

Las variables de la base de datos son:

- **ID:** Identificador del cliente.
- **LIMIT_BAL:** Límite de crédito.
- **SEX:** Sexo (1 = masculino, 2 = femenino).
- **EDUCATION:** Educación (1=grad, 2=univ, 3=secundaria, 4=otros, 5=desconocido, 6=desconocido).
- **MARRIAGE:** Estado civil (1=casado, 2=soltero, 3=otros).
- **AGE:** Edad del cliente.
- **PAY_0 a PAY_6:** Historial de pagos de los últimos 6 meses (0 = al día, valores positivos = retrasos).
- **BILL_AMT1 a BILL_AMT:** Monto de facturas de los últimos 6 meses.
- **PAY_AMT1 a PAY_AMT6:** Monto de pagos realizados en los últimos 6 meses.
- **default_payment:** Variable objetivo: 0 = no entró en default, 1 = entró en default.

El dataset tiene 30.000 filas y 25 columnas.

13 columnas son float64 (montos de crédito, facturación y pagos) y 12 columnas son int64 (variables categóricas y default_payment).

La media de default_payment es 0.2212, es decir, un 22,1% de clientes en default en el dataset.

En variables como PAY_0, PAY_2, … PAY_6 (historial de pagos), los valores van de -2 a 8, donde: -2, -1, 0 suelen indicar pagos al día o anticipados. Mientras que valores positivos indican meses de atraso.

LIMPIEZA DE DATOS:

La variable **EDUCATION** debería tener las categorías:

- 1 = graduate school
- 2 = university
- 3 = high school

Pero en el dataset aparecen valores raros: 0, 5, 6. Estos no tienen sentido según la definición de las variables del dataset, así que se reemplazan por 4, que se interpreta como “otros/otros estudios”.

La variable **MARRIAGE** debería ser:

- 1 = casado
- 2 = soltero

Aparece un 0 que no tiene sentido. Se reemplaza por 3, interpretado como “otros”.

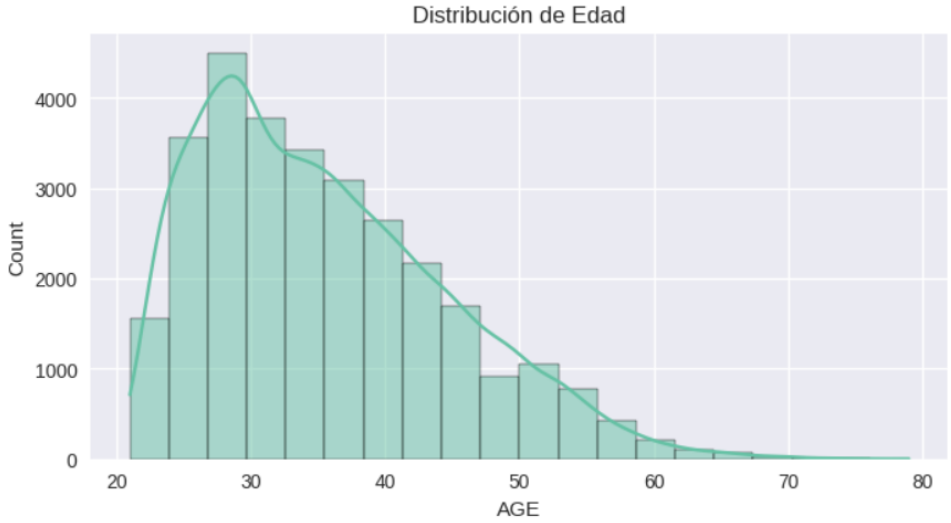
Además, hemos comprobado los valores nulos, pero vemos que no existen ninguna de las variables, por lo que no es necesario aplicar técnicas de imputación.

También vamos a eliminar variables que son poco relevantes como la variable ID.

ANÁLISIS INDIVIDUAL DE VARIABLES

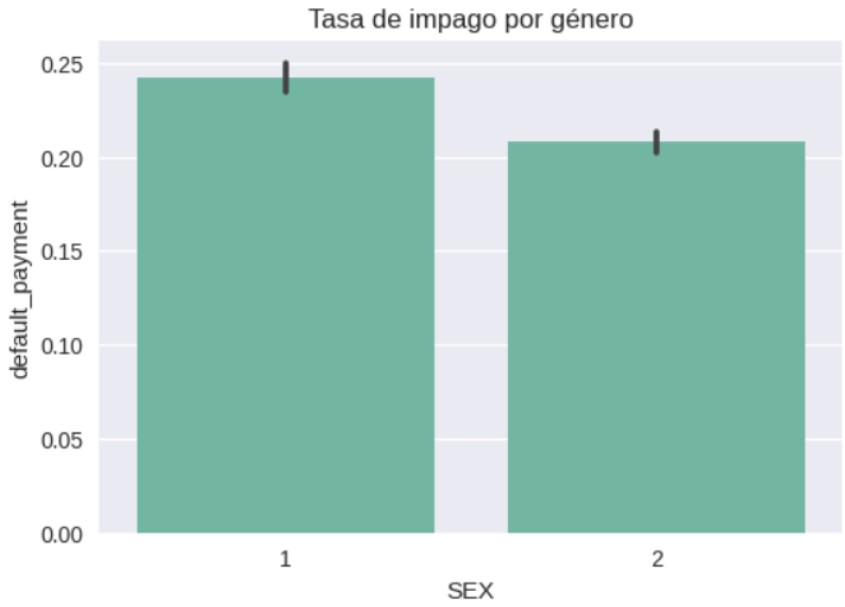
HISTOGRAMA DE EDAD

La mayoría de los clientes son adultos jóvenes y de mediana edad, lo que puede ser relevante para segmentación o análisis de riesgo.



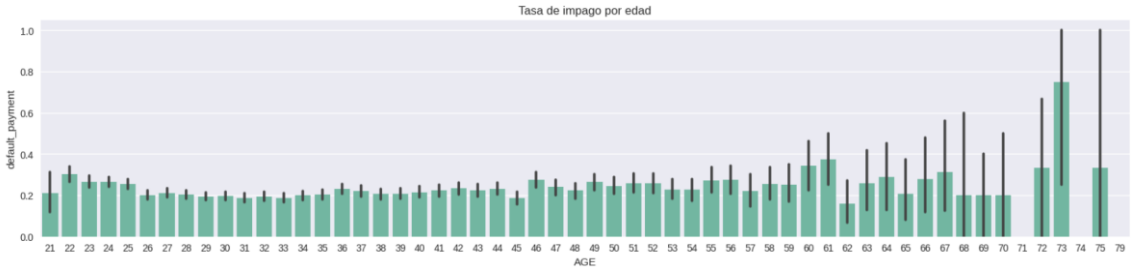
PORCENTAJE DE IMPAGOS POR GÉNERO

Los hombres tienen más tendencia a entrar en impago que las mujeres.



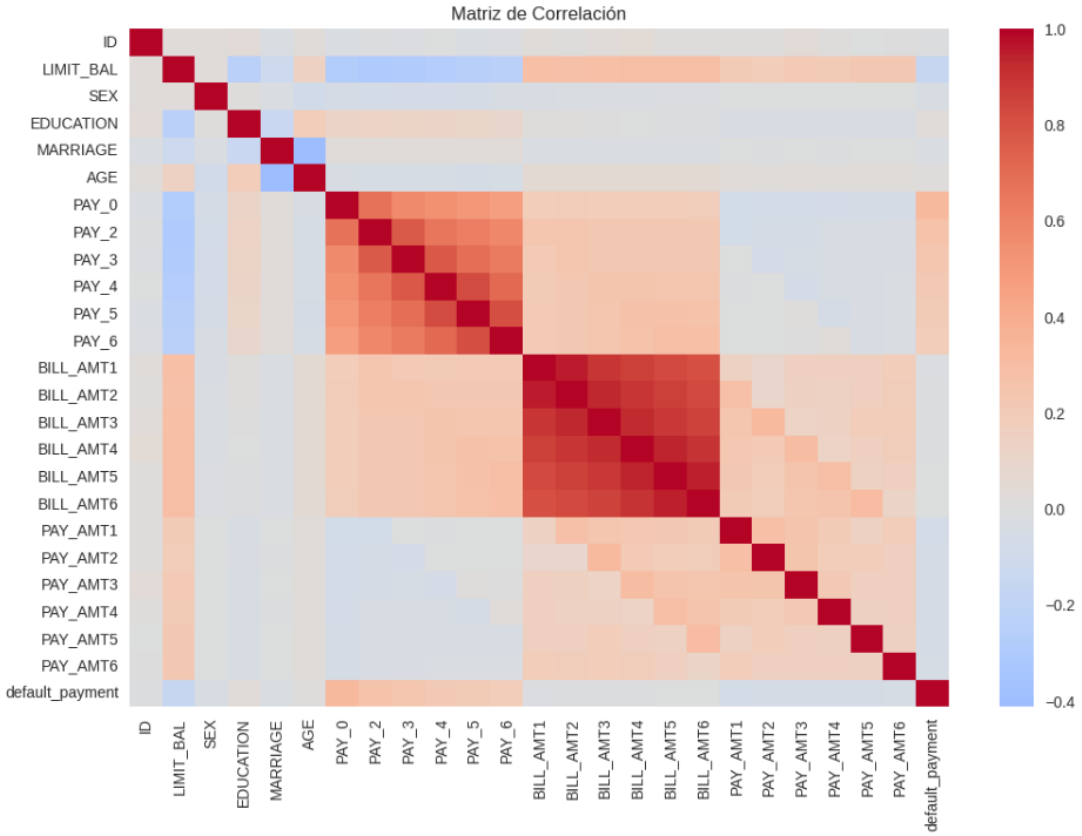
PORCENTAJE DE IMPAGOS POR SEXO

Clientes muy jóvenes (21–24 años) y mayores de 60 años muestran una mayor variabilidad y algunas tasas de impago más altas. Las edades más avanzadas (70–79) muestran tasas de impago más elevadas, aunque la muestra es más pequeña.



CORRELACIÓN

Vamos a ver la relación lineal entre las variables numéricas.



En este caso se observa que los históricos de pagos (PAY_0 a PAY_6) y los importes facturados (BILL_AMT1 a BILL_AMT6) están fuertemente correlacionados entre sí, lo cual es esperable al tratarse de series mensuales consecutivas. Además, la variable objetivo default_payment muestra mayor correlación con el estado de pago más reciente (PAY_0), lo que confirma que el historial de pagos previos es un buen predictor del riesgo de impago. En cambio, variables demográficas como edad, sexo o estado civil presentan baja correlación con el impago.

MODELADO

Para empezar a preparar los modelos, vamos a preparar los datos. Para ellos dividimos los datos entre las variables x (todas menos default_payment) y nuestra variable objetivo, y (default_payment). También tenemos que estandarizar los datos para modelos lineales.

LOGISTIC REGRESSION

```
=== Logistic Regression ===
              precision    recall  f1-score   support

           0       0.82      0.97      0.89      7009
           1       0.69      0.24      0.35      1991

    accuracy                           0.81      9000
   macro avg       0.76      0.60      0.62      9000
weighted avg       0.79      0.81      0.77      9000

ROC-AUC: 0.715023533995432
```

El modelo de regresión logística acierta bien los clientes que pagan, pero tiene bajo recall en impagos (24%). Es un buen modelo base, interpretable y rápido.

RANDOM FOREST

```
=== Random Forest ===
              precision    recall  f1-score   support

           0       0.84      0.94      0.89      7009
           1       0.65      0.37      0.47      1991

    accuracy                           0.82      9000
   macro avg       0.74      0.66      0.68      9000
weighted avg       0.80      0.82      0.80      9000

ROC-AUC: 0.758976637556979
```

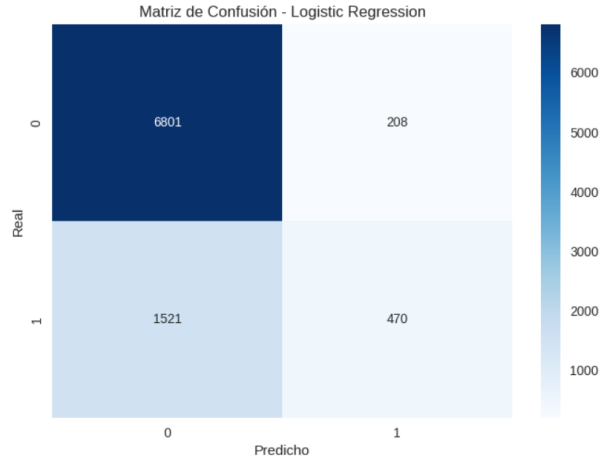
El modelo random forest mejora ligeramente respecto a la regresión logística. Detecta más impagos (recall 37%), aunque sigue costándole identificar bien la clase minoritaria.

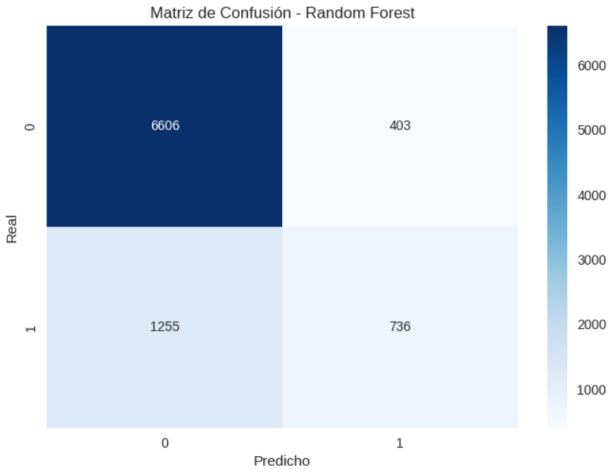
XG BOOST

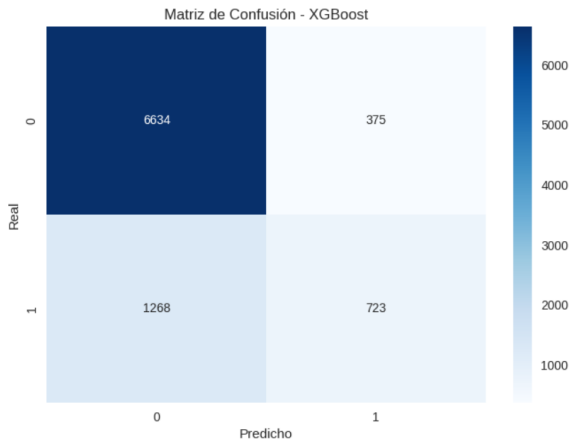
```
=== XGBoost ===
              precision    recall  f1-score   support

           0       0.84      0.95      0.89      7009
           1       0.66      0.36      0.47      1991

    accuracy                           0.82      9000
   macro avg       0.75      0.65      0.68      9000
weighted avg       0.80      0.82      0.80      9000

ROC-AUC: 0.7698392588305243
```

Es el mejor modelo en términos globales. Consigue el mayor poder de discriminación entre pagadores e impagadores, aunque el recall de impagos sigue siendo bajo (36%).

MATRIZ CONFUSIÓN



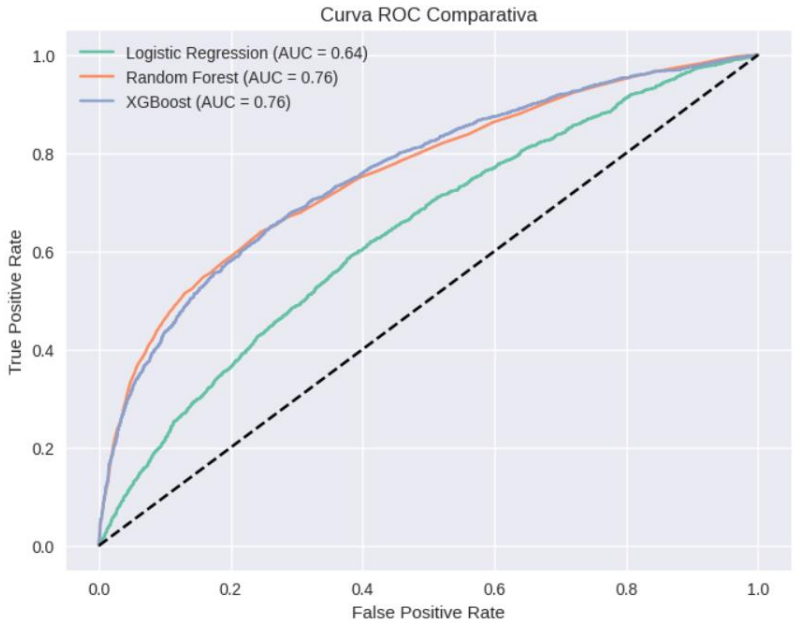




La Regresión Logística muestra un buen desempeño identificando a los clientes que no impagan (muchos verdaderos negativos), pero tiene dificultades para detectar correctamente a los que sí incumplen, lo que se refleja en un alto número de falsos negativos y un recall bajo. Por su parte, tanto el Random Forest como XGBoost logran un mejor equilibrio: aunque siguen clasificando correctamente a la mayoría de los no impagos, aumentan la capacidad de identificar clientes morosos (mayor número de verdaderos positivos), reduciendo los falsos

negativos en comparación con la regresión logística. Entre ellos, XGBoost ofrece el mejor rendimiento global, mostrando un buen compromiso entre precisión general y detección de impagos, lo que lo convierte en el modelo más adecuado para este problema.

CURVA ROC COMPARATIVA

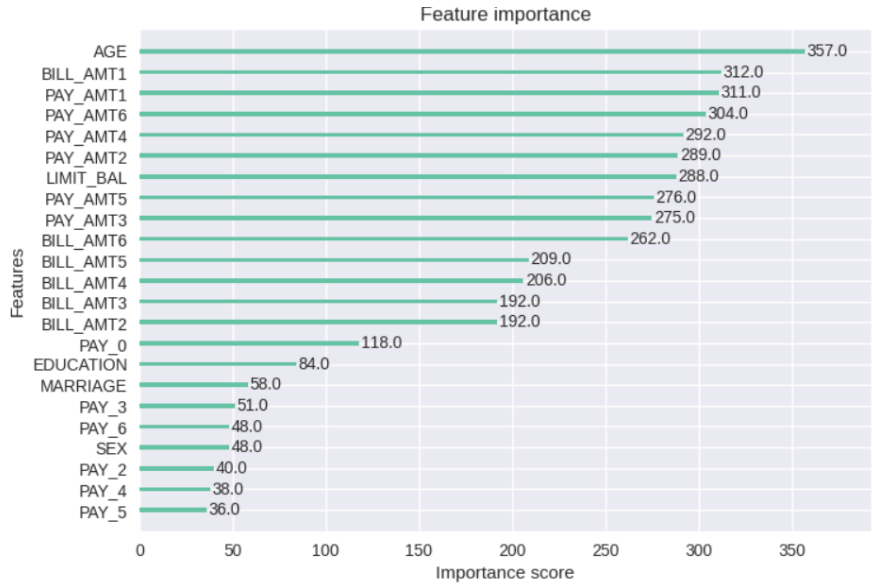


El gráfico muestra la curva ROC, que refleja la capacidad de los modelos para diferenciar entre clientes que pagan y los que no. Cuanto más se acerque la curva a la esquina superior izquierda, mejor es la capacidad de predicción. El área bajo la curva (AUC) indica el rendimiento global del modelo: valores más próximos a 1 significan mayor precisión en la clasificación.

Nos quedamos con el modelo XGBoost por ser el que mejor rendimiento a tenido a lo largo del análisis.

Por tanto, procedemos a guardar el modelo para que después se pueda reutilizar para predecir.

Por último, vemos qué factores son los que más afectan a las predicciones:



El modelo XGBoost muestra que las variables más importantes para sus predicciones son la edad (AGE), los montos de las facturas recientes (BILL_AMT1) y los pagos realizados en distintos meses (PAY_AMT1, PAY_AMT6, PAY_AMT2). Esto indica que tanto la edad como el historial de pagos recientes son clave para predecir el comportamiento.

CONCLUSIÓN FINAL

El conjunto de datos UCI Credit Card constituye una fuente de información relevante para el análisis del riesgo de impago en clientes con créditos financieros.

Tras un proceso de limpieza y depuración de los datos, se procedió a la construcción de modelos de Machine Learning, empleando técnicas de regresión logística, Random Forest y XGBoost.

Los resultados obtenidos evidencian que los modelos Random Forest y XGBoost alcanzaron el mejor desempeño, con valores de AUC en torno a 0,76–0,77, lo que refleja una capacidad razonable para discriminar entre clientes solventes e insolventes.

No obstante, la recall asociada a la clase de impago resultó baja en todos los modelos evaluados, lo que implica que, si bien los algoritmos logran predecir con mayor eficacia a los clientes cumplidores, presentan dificultades para identificar adecuadamente a los que incurrirán en incumplimiento.

En un escenario aplicado, podrían implementarse estrategias adicionales como el reajuste de clases, la optimización de hiperparámetros o la incorporación de nuevas variables explicativas, con el fin de mejorar la capacidad predictiva sobre la clase minoritaria y, en particular, incrementar la recall en los casos de impago.