CODER HOUSE

PROYECTO FINAL DE DATA SCIENCE: **INCUMPLIMIENTO CREDITICIO**

Profesor: David Francisco Bustos Usta

Tutor: Gianluca Peretti

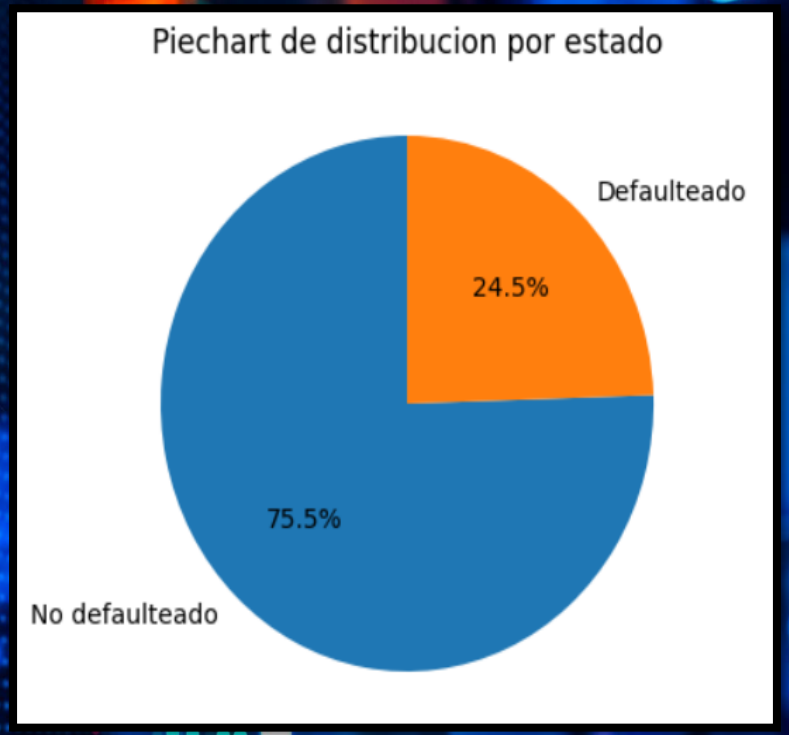
Grupo: Ariana Diaz y Federico Martucci

Fecha: 26/06/2022

TABLA DE CONTENIDOS

OBJETIVO

Determinar si un cliente con determinadas características pagará o no el préstamo.



Dataset desbalanceado

CONTENIDO DATASET

Filas: 148.670
Columnas: 34

VARIABLES CATEGÓRICAS

- Límite del préstamo (con o sin)
- Género del cliente
- Pre aprobado (si o no)
- Tipo de préstamo
- Lump sum payment(si o no)
- Rango etario del cliente
- Tipo de crédito del co aplicante
- Negocio/ comercial o particular
- Amortización negativa

Fuente: https://www.kaggle.com/datasets/yasserh/loan-default-dataset

CONTENIDO DATASET

Filas: 148.670
Columnas: 34

VARIABLES NUMÉRICAS

- Tasa de interés
- Otros cargos del crédito
- Valor de la propiedad
- Tipo de préstamo
- Ingresos
- Puntaje crediticio
- Plazo del préstamo
- ID
- LTV

Fuente: https://www.kaggle.com/datasets/yasserh/loan-default-dataset

CONTENIDO DATASET

Filas: 148.670
Columnas: 34

VARIABLE RESPUESTA

Status

- 0: Pagará
- 1: No pagará

Fuente: https://www.kaggle.com/datasets/yasserh/loan-default-dataset

DATA WRANGLING

Valores faltantes	Se completan con la media
Variables categóricas a numéricas	Label Encoder
Outliers	Eliminar Q_1 y Q_4
Normalización de datos	Yeo-johnson
Selección de variables	SelectKBest

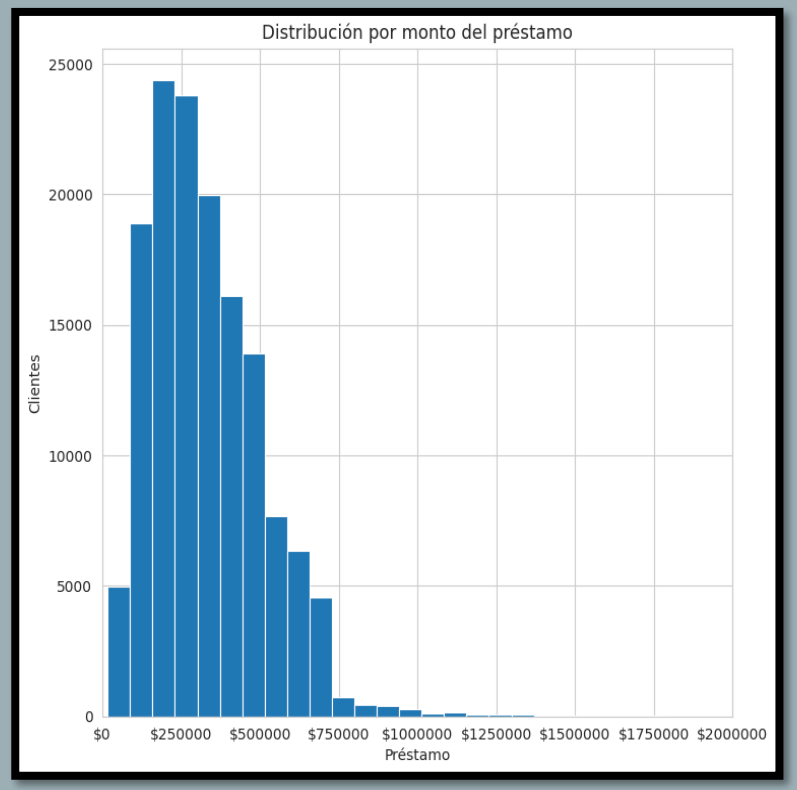
VARIABLES DESTACADAS

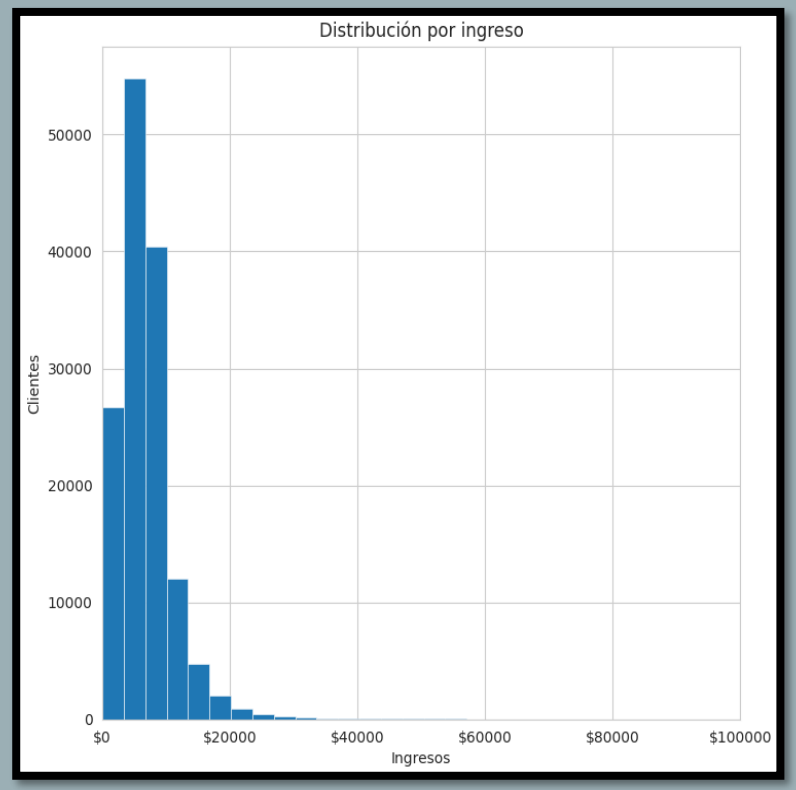
DATASET LIMPIO

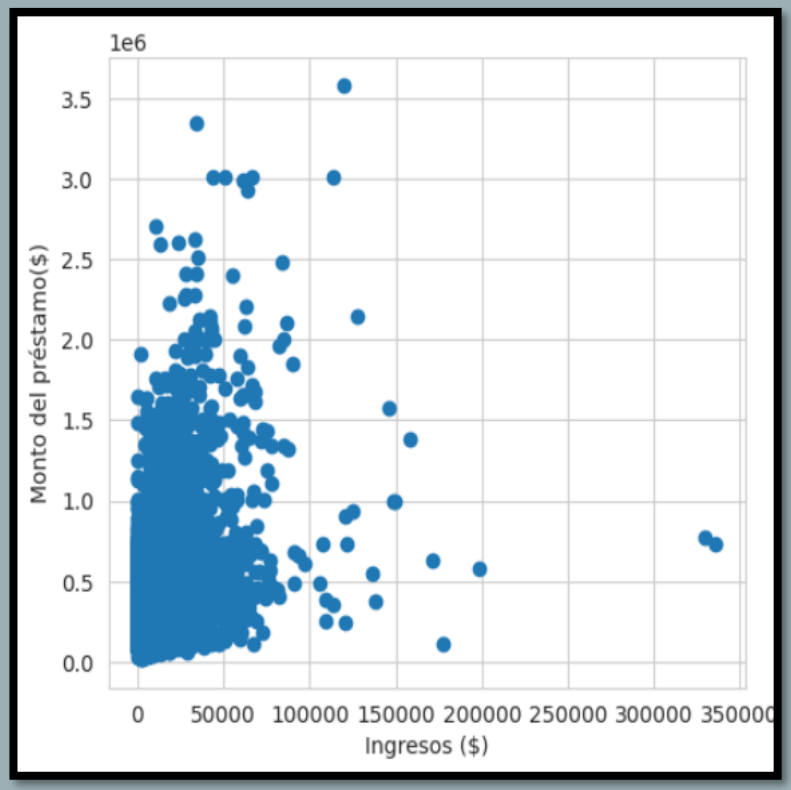
Filas: 143.162
Columnas: 5

Otros cargos	• Otros cargos del préstamo, fuera de los intereses. • En el EDA se explicará porque no se puede tener en cuenta.
Ingresos	• Ingresos del cliente que toma el crédito.
Pago global	• Pago único, más grande de lo normal, que se hace al final del plazo del préstamo. • Puede ser si o no.
Tipo de crédito del co aplicante	• Un cofirmante, por lo general un miembro de la familia, ayuda para que le aprueben un préstamo a un prestatario, al aceptar pagarlo si este no lo hace. • Puede ser CIB o EXP
Tipo de crédito	• Puede ser EXP, CIB, CRIF y EQUI.

EXPLORATORY DATA ANALYSIS

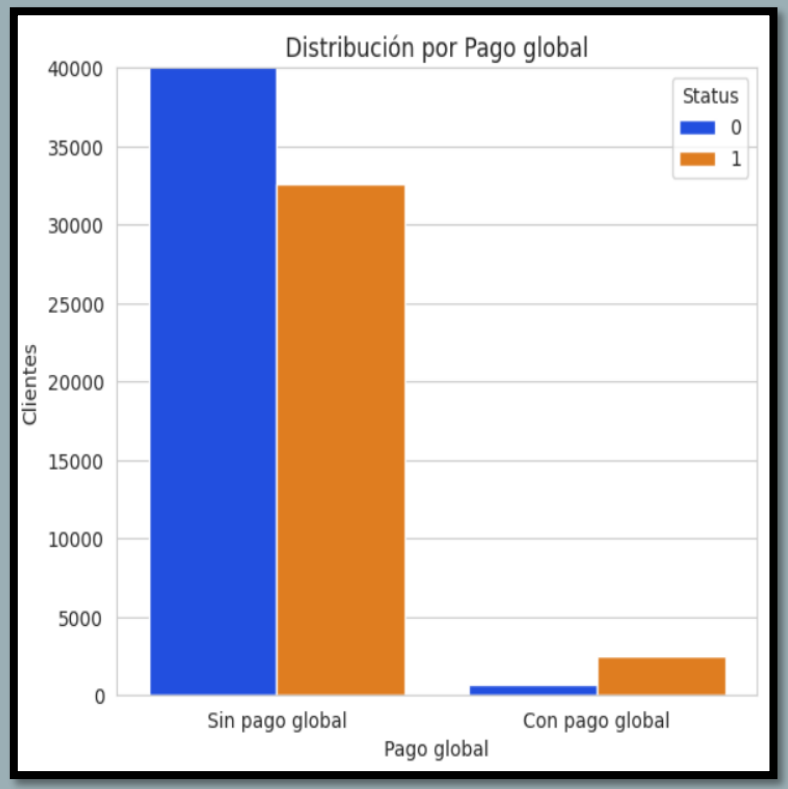


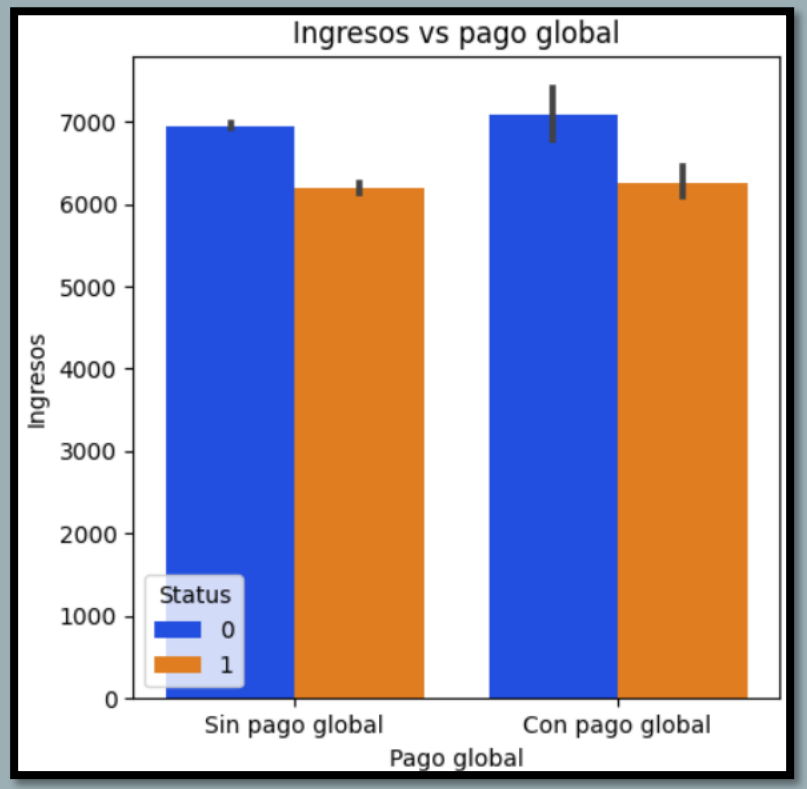




- Mucha gente con pocos ingresos, poca gente con muchos ingresos.
- La gente de menores ingresos pide más préstamos. El monto de estos no es alto.

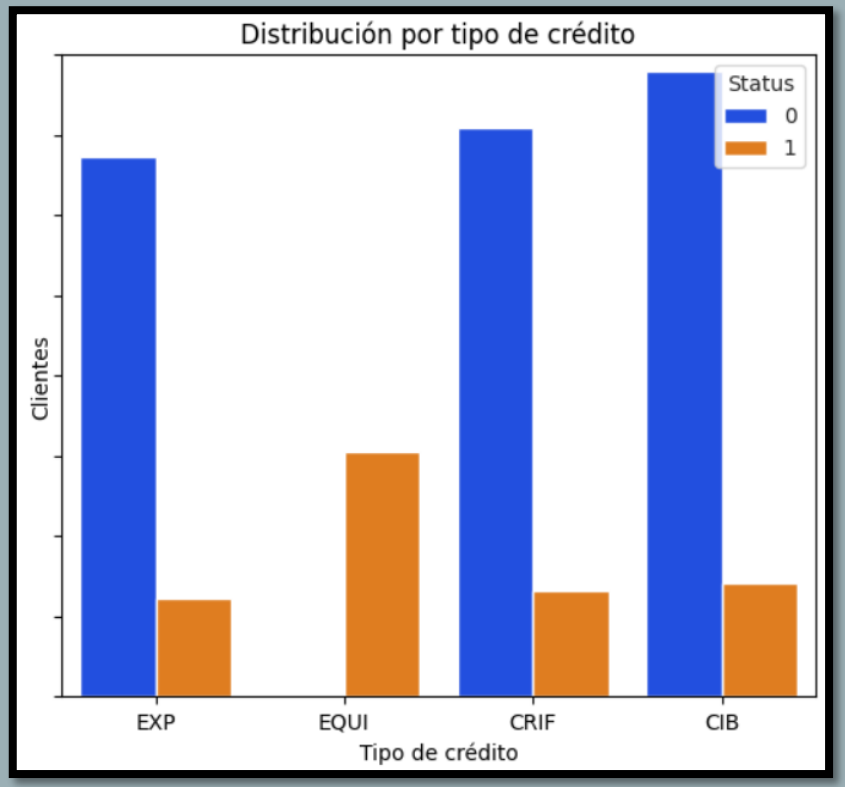
EXPLORATORY DATA ANALYSIS

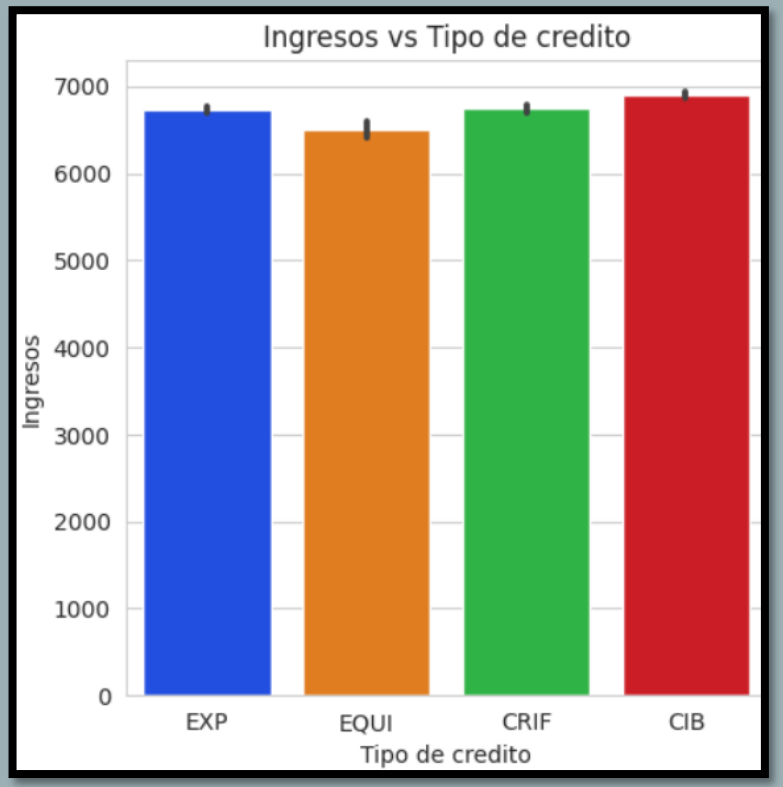




- La mayoría de los créditos con pago global no se pagarán, aunque son pocos casos.
- Tanto en los créditos con y sin pago global, el default del mismo está asociado a salarios mas bajos.

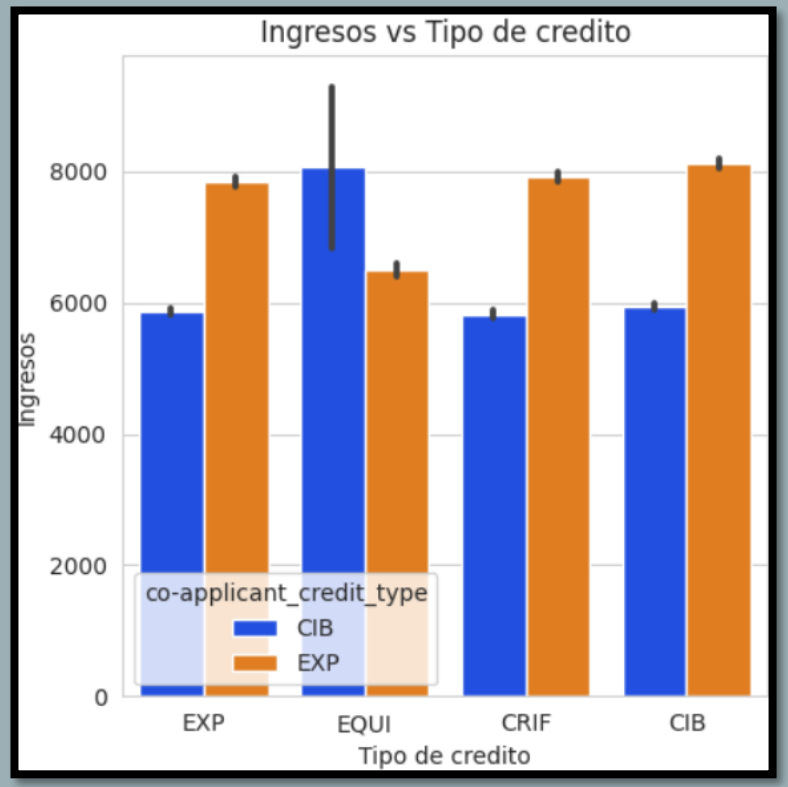
EXPLORATORY DATA ANALYSIS

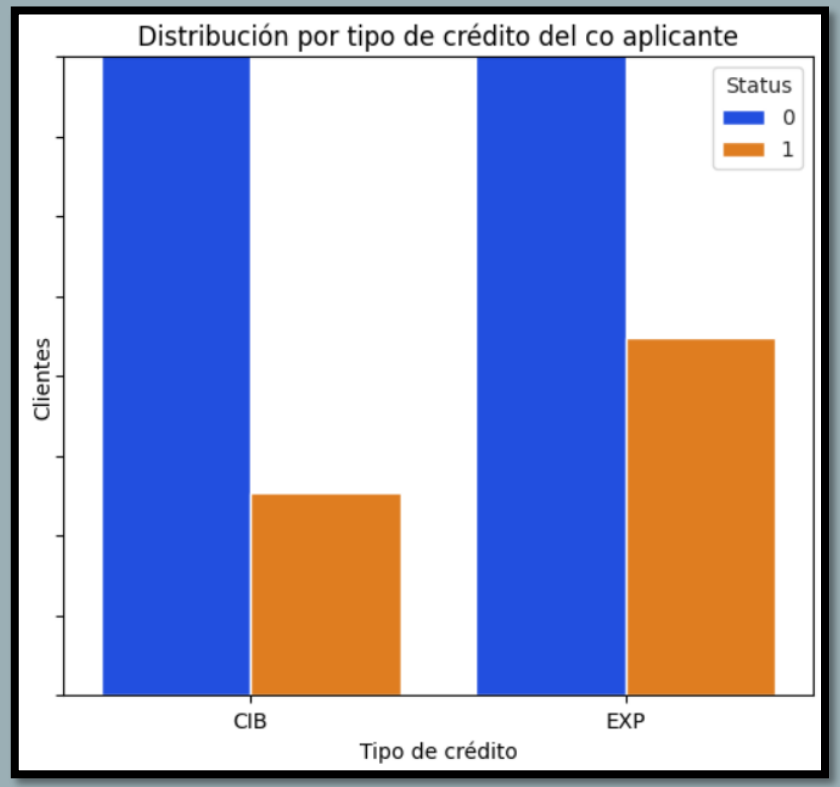




- La gran mayoría de los créditos tipo EQUI no se pagarán, y además, están asociados a ingresos mas bajos.

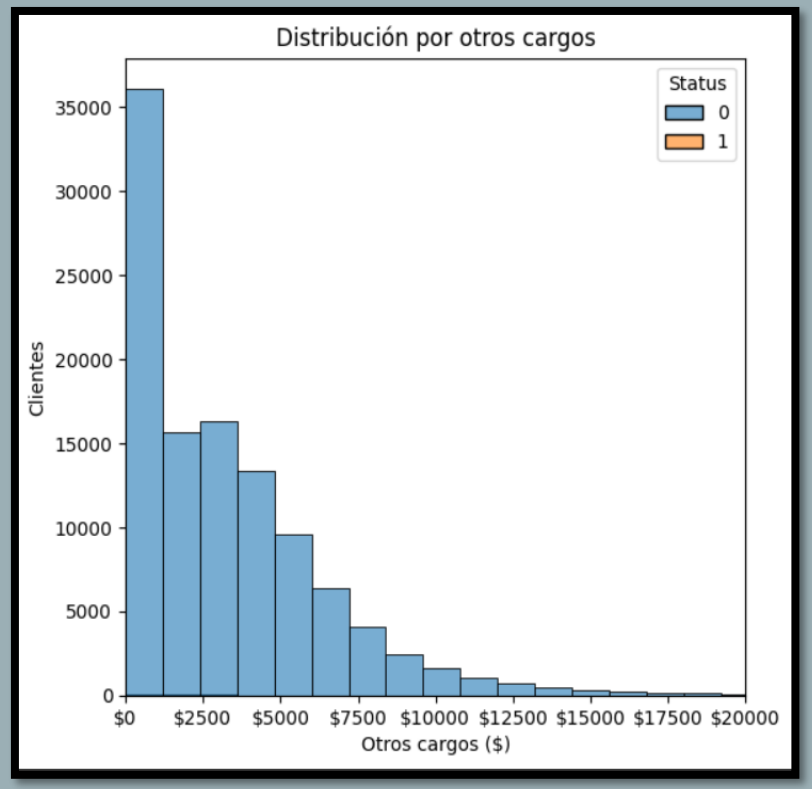
EXPLORATORY DATA ANALYSIS

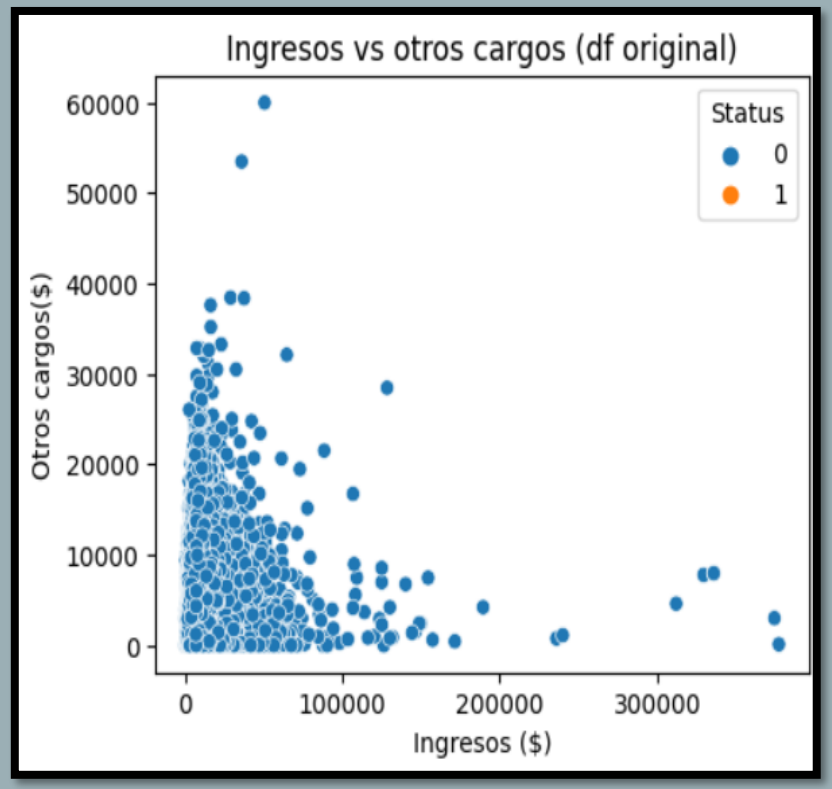


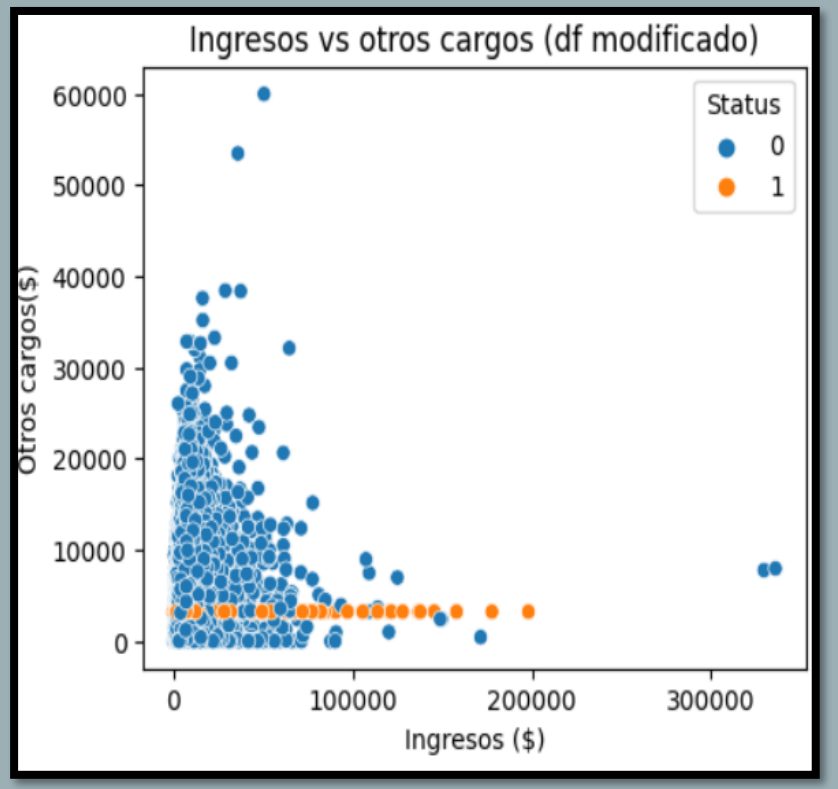


- En cuanto a la relación del tipo de crédito con el tipo de crédito del coaplicante, en todos los casos, excepto EQUI, los clientes con mayores ingresos tienen un coaplicante tipo EXP.
- Los créditos de coaplicantes tipo EXP tienen mas probabilidades de no pagar.

OTROS CARGOS DEL CRÉDITO







- Otros cargos no puede ser considerado, dado que cuando target es positiva, los datos de esta variable son *null*.
- Únicamente cuando se completan los *nulls* del dataset con la media comienza a aparecer la variable target como positiva. Para no generar ruido se decide no considerarla.

CODER HOUSE

MODELOS DE CLASIFICACIÓN

La variable más importante es la **Sensibilidad (recall)**.

Nos indica la capacidad de dar con casos positivos, es decir, clientes que no pagarán el préstamo tomado.

La segunda variable a tener en cuenta es la **Precisión**.

Resume el rendimiento de un modelo de clasificación cuando se tienen dos clases con tamaño desigual.

PARA EVALUAR AMBOS DE MANERA CONJUNTA SE USA F1 SCORE.

CODER HOUSE

MODELOS DE CLASIFICACIÓN

Regresión Logística	• F1 = 0,46 ✗
Árbol de decisión	• F1 = 0,61 ✓
KNN	• F1 = 0,61 ✓
Random Forest	• F1 = 0,61 ✓
SCV	• F1 = 0,62 • Tarda mucho en correr ✗

CODER HOUSE

VALIDACION DE MODELOS CON STRATIFIED K FOLD

Arbol de decisión	• F1 promedio: 0,609 • Sensibilidad promedio: 0,444 ✗
Random Forest	• F1 promedio: 0,611 • Sensibilidad promedio: 0,534 ✓
KNN	• F1 promedio: 0,597 • Sensibilidad promedio: 0,469 ✓

5 iteraciones

CODER HOUSE

HYPERTUNING DE PARAMETROS

Random Forest	
$F1_{prom}$=0,623	$Sensibilidad_{prom}$=0,531 ✓
KNN	
$F1_{prom=}$0,626	$Sensibilidad_{prom}$=0,469 ✗

RandomizedSearchCV

MODELO ELEGIDO: RANDOM FOREST CLASSIFIER

Parámetros:

n_estimators: 1600

min_samples_split: 10

min_samples_leaf: 1

max_features: 'sqrt'

max_depth: 10

bootstrap: False

CONCLUSIONES FINALES

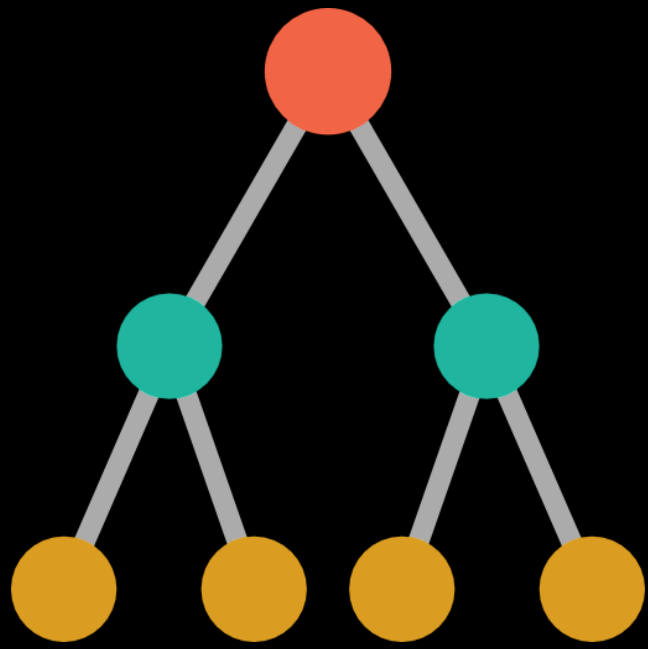
RANDOM FOREST CLASSIFIER

F1=0,62

SENSIBLIDAD=0,46

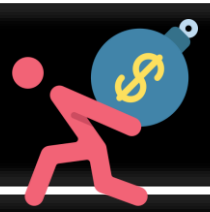
ACCURACY=0,86

PRECISION=0,95



CONCLUSIONES FINALES

25% DE LOS CLIENTES NO ABONARÁ EL CRÉDITO



CLIENTES CON MENORES INGRESOS SON MAS PROPENSOS A NO ABONAR EL CRÉDITO

MODELO QUE MEJOR SE ADAPTA: RANDOM FOREST CLASSIFIER, CON 86% DE ACCURACY

FUTURAS LINEAS

¿CÓMO COMPLEMENTAR EL PROYECTO?

- ANÁLISIS DE LA VARIABLE 'OTROS CARGOS' EN RELACION CON TARGET
- ANÁLISIS SOBRE POSIBLES DIFERENCIAS DE GÉNERO A LA HORA DE TOMAR CRÉDITOS
- PROBAR MODELO SCV
- EMPLEAR OTROS MÉTODOS PARA REDUCIR AUN MÁS LOS OUTLIERS

CODER HOUSE

GRACIAS

Profesor: David Francisco Bustos Usta

Tutor: Gianluca Peretti

Grupo: Ariana Diaz y Federico Martucci

Fecha: 26/06/2022