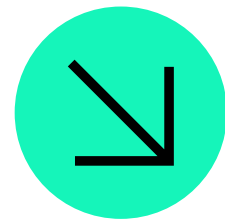


¿A quiénes les podemos prestar dinero?

El intento de un algoritmo
de clasificación efectivo





Contiene alrededor de 310,000 filas y 122 columnas, de las cuales, 1 nos indicaba si las personas habían hecho default o no, siendo 1 default y 0 no default

Nuestro set de datos

En el caso de los préstamos bancarios, “default” se refiere a la situación en la que el usuario deja de pagar un préstamo dado



Objetivos

- Generar un algoritmo clasificador para definir qué personas entrarán en “default” o no
- Definir qué algoritmo y mecanismos de procesamiento de datos son los mejores para lograr un algoritmo eficiente



Variables en el set de datos

- SK_ID_CURR
- TARGET
- NAME_CONTRACT_TYPE
- CODE_GENDER
- FLAG_OWN_CAR
- FLAG_OWN_REALTY
- CNT_CHILDREN
- AMT_INCOME_TOTAL
- AMT_CREDIT
- AMT_ANNUITY
- AMT_GOODS_PRICE
- NAME_TYPE_SUITE
- NAME_INCOME_TYPE
- NAME_EDUCATION_TYPE
- NAME_FAMILY_STATUS
- NAME_HOUSING_TYPE
- REGION_POPULATION_RELATIVE
- DAYS_BIRTH
- DAYS_EMPLOYED
- DAYS_REGISTRATION
- DAYS_ID_PUBLISH
- OWN_CAR_AGE
- FLAG_MOBIL
- FLAG_EMP_PHONE
- FLAG_WORK_PHONE
- FLAG_CONT_MOBILE
- FLAG_PHONE

- FLAG_EMAIL
- OCCUPATION_TYPE
- CNT_FAM_MEMBERS
- REGION_RATING_CLIENT
- REGION_RATING_CLIENT_W_CITY
- WEEKDAY_APPR_PROCESS_START
- HOUR_APPR_PROCESS_START
- REG_REGION_NOT_LIVE_REGION
- REG_REGION_NOT_WORK_REGION
- LIVE_REGION_NOT_WORK_REGION
- REG_CITY_NOT_LIVE_CITY
- REG_CITY_NOT_WORK_CITY
- LIVE_CITY_NOT_WORK_CITY
- ORGANIZATION_TYPE
- EXT_SOURCE_1
- EXT_SOURCE_2
- EXT_SOURCE_3
- APARTMENTS_AVG
- BASEMENTAREA_AVG
- YEARS_BEGINEXPLUATATION_AVG
- YEARS_BUILD_AVG
- COMMONAREA_AVG
- ELEVATORS_AVG
- ENTRANCES_AVG
- FLOORSMAX_AVG
- FLOORSMIN_AVG
- LANDAREA_AVG

- LIVINGAPARTMENTS_AVG
- LIVINGAREA_AVG
- NONLIVINGAPARTMENTS_AVG
- NONLIVINGAREA_AVG
- APARTMENTS_MODE
- BASEMENTAREA_MODE
- YEARS_BEGINEXPLUATATION_MODE
- YEARS_BUILD_MODE
- COMMONAREA_MODE
- ELEVATORS_MODE
- ENTRANCES_MODE
- FLOORSMAX_MODE
- FLOORSMIN_MODE
- LANDAREA_MODE
- LIVINGAPARTMENTS_MODE
- LIVINGAREA_MODE
- NONLIVINGAPARTMENTS_MODE
- NONLIVINGAREA_MODE
- APARTMENTS_MEDI
- BASEMENTAREA_MEDI
- YEARS_BEGINEXPLUATATION_MEDI
- YEARS_BUILD_MEDI
- COMMONAREA_MEDI
- ELEVATORS_MEDI
- ENTRANCES_MEDI
- FLOORSMAX_MEDI
- FLOORSMIN_MEDI

1

Cualitativas

2

Cuantitativas

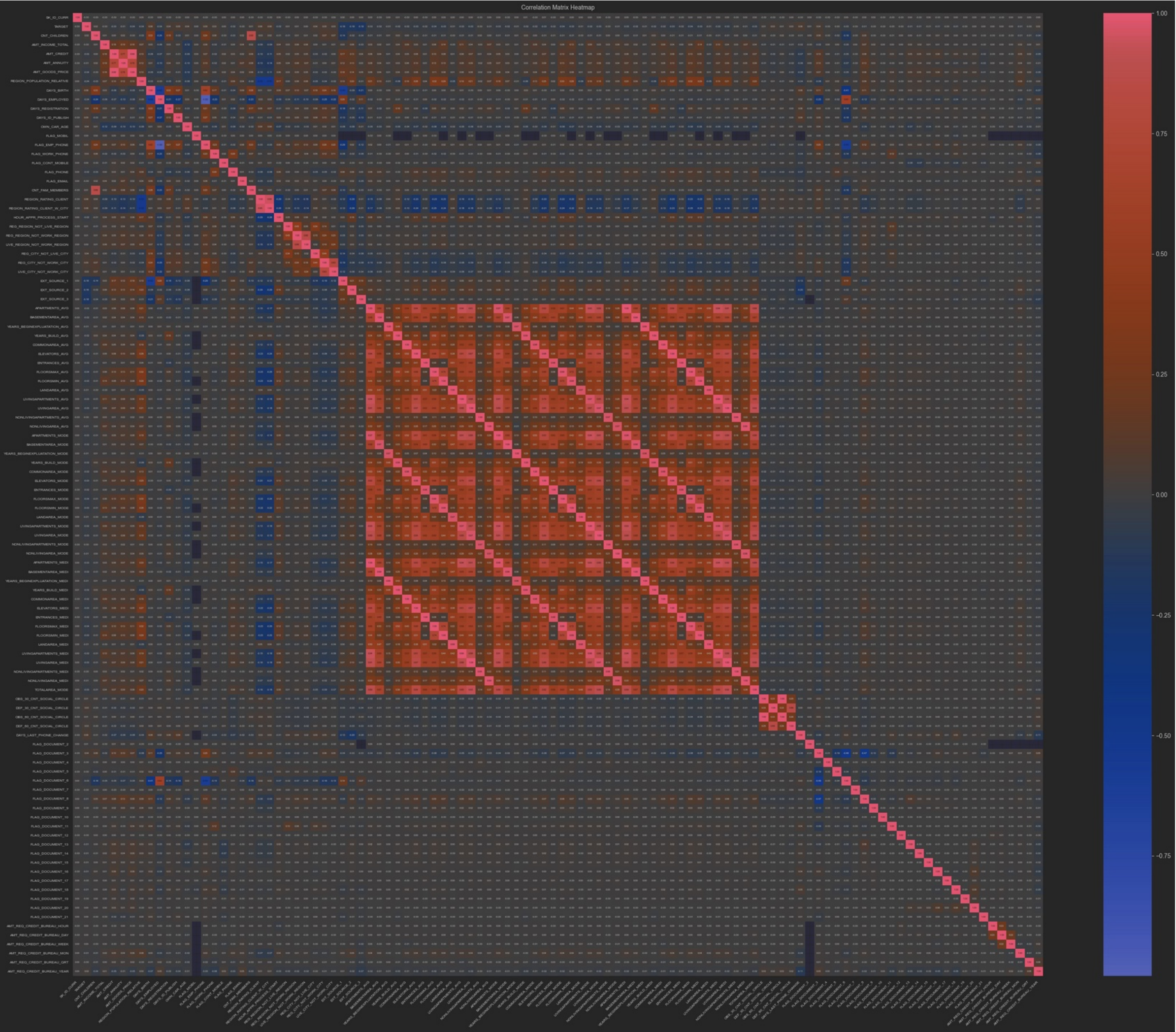
3

Variable a predecir

- LANDAREA_MEDI
- LIVINGAPARTMENTS_MEDI
- LIVINGAREA_MEDI
- NONLIVINGAPARTMENTS_MEDI
- NONLIVINGAREA_MEDI
- FONDKAPREMONT_MODE
- HOUSETYPE_MODE
- TOTALAREA_MODE
- WALLSMATERIAL_MODE
- EMERGENCYSTATE_MODE
- OBS_30_CNT_SOCIAL_CIRCLE
- DEF_30_CNT_SOCIAL_CIRCLE
- OBS_60_CNT_SOCIAL_CIRCLE
- DEF_60_CNT_SOCIAL_CIRCLE
- DAYS_LAST_PHONE_CHANGE
- FLAG_DOCUMENT_2
- FLAG_DOCUMENT_3
- FLAG_DOCUMENT_4
- FLAG_DOCUMENT_5
- FLAG_DOCUMENT_6
- FLAG_DOCUMENT_7
- FLAG_DOCUMENT_8
- FLAG_DOCUMENT_9
- 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_20
- FLAG_DOCUMENT_21
- AMT_REQ_CREDIT_BUREAU_HOUR
- AMT_REQ_CREDIT_BUREAU_DAY
- AMT_REQ_CREDIT_BUREAU_WEEK
- AMT_REQ_CREDIT_BUREAU_MON
- AMT_REQ_CREDIT_BUREAU_QRT
- AMT_REQ_CREDIT_BUREAU_YEAR

Correlación





Cambios iniciales

1. Algunas fechas se encontraban en días “negativos” por lo que aplicamos valor absoluto
2. Existían columnas que podíamos sumar para obtener una sola y reducir la cantidad de variables (ej. documento 1, documento 2)
3. Teníamos demasiados NA para algunas variables, por lo que rellenamos estos utilizando clusters

1

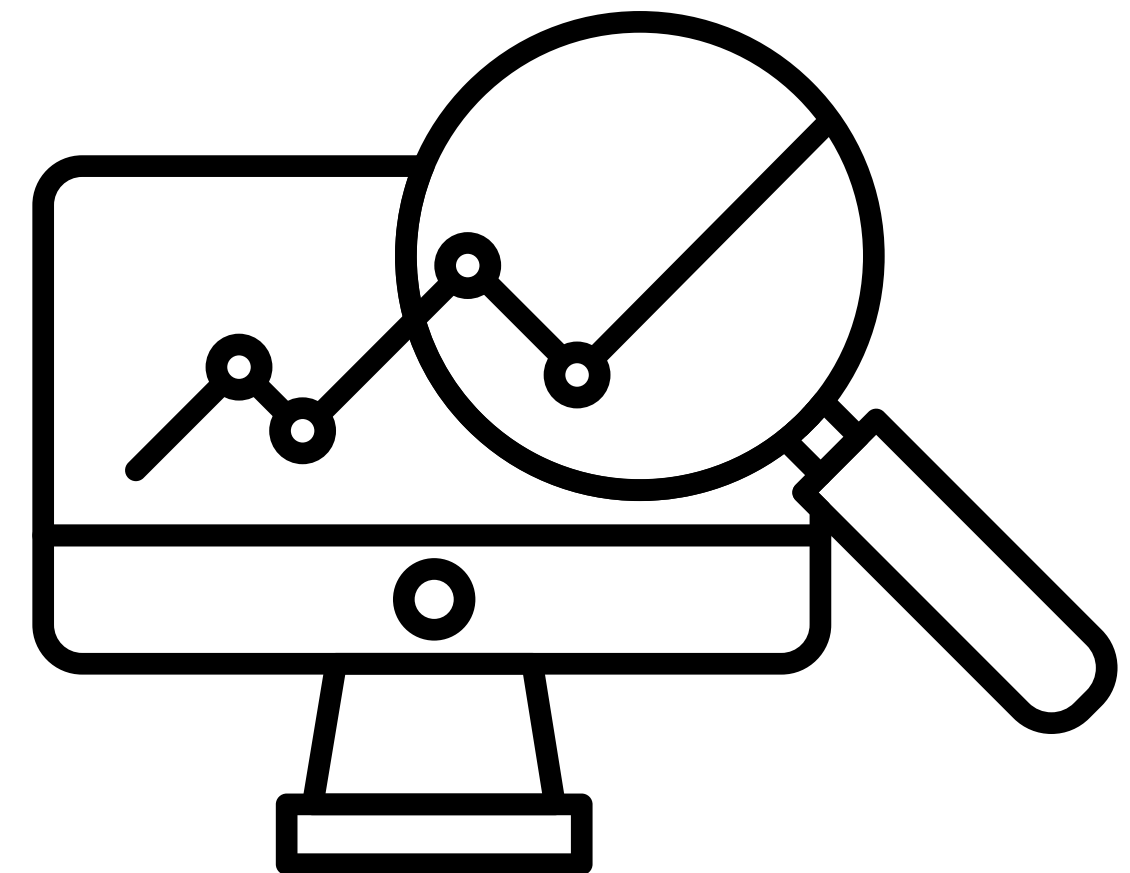
Transformar en positivos

2

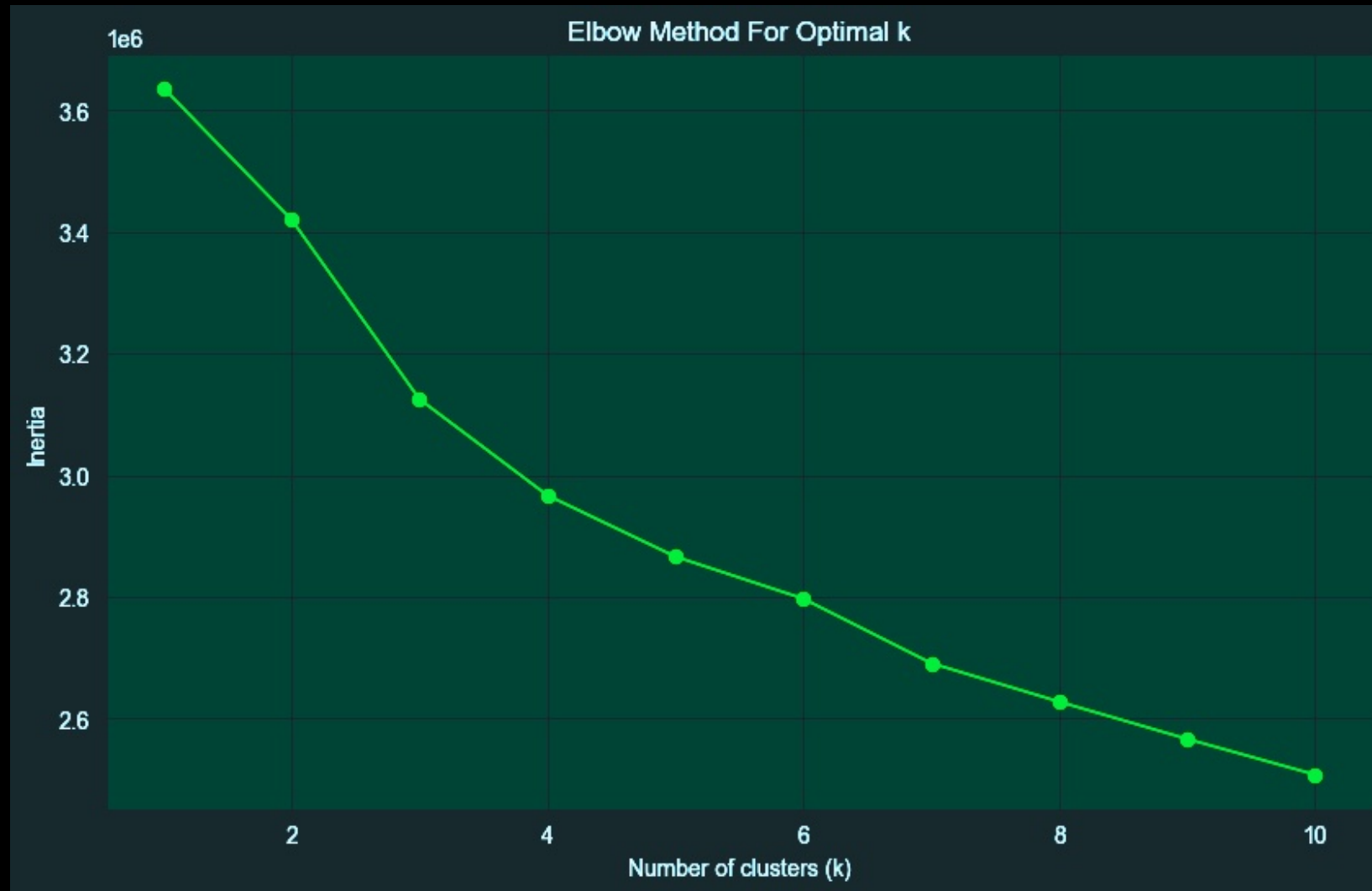
Juntar columnas

3

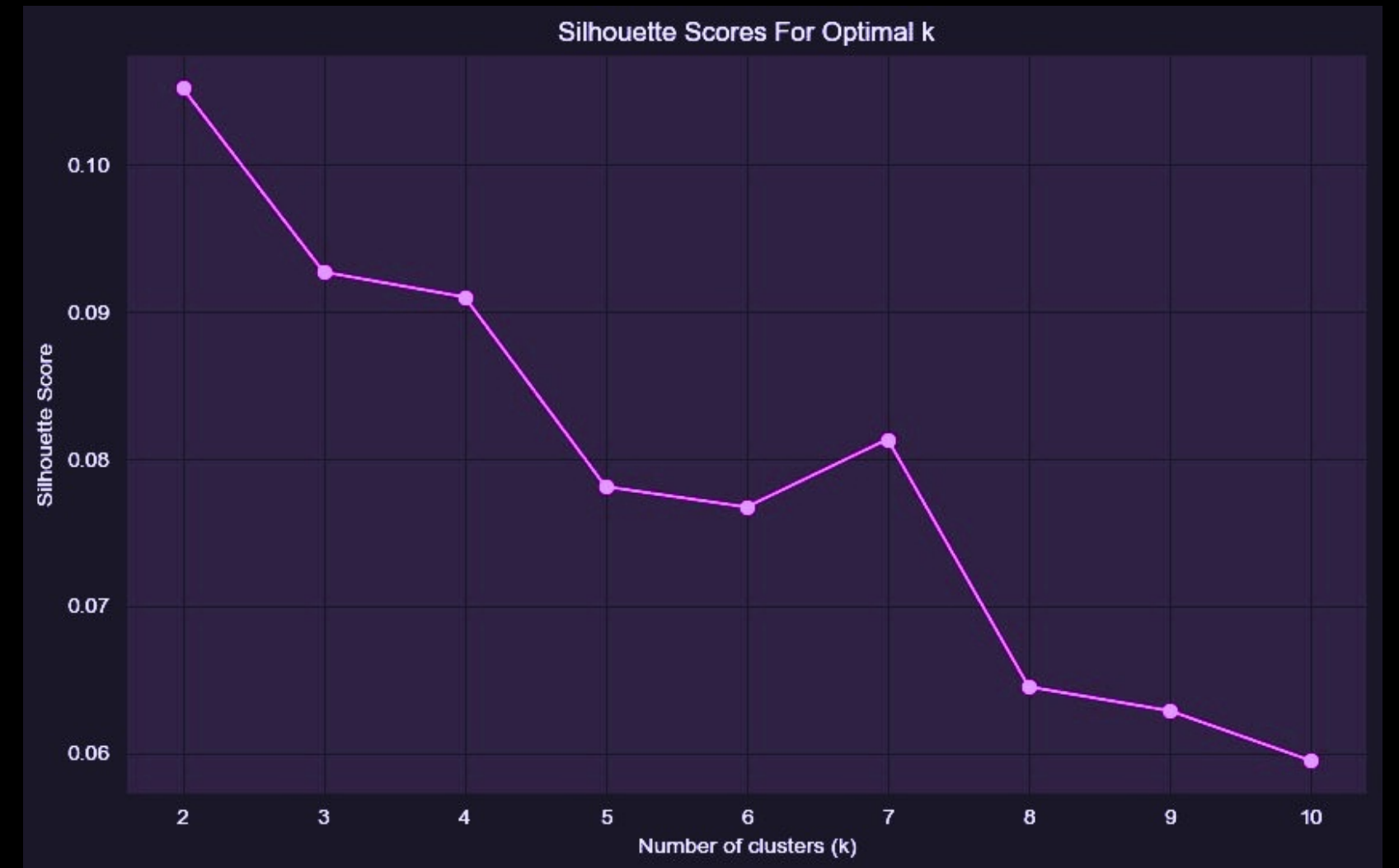
Rellenar NA



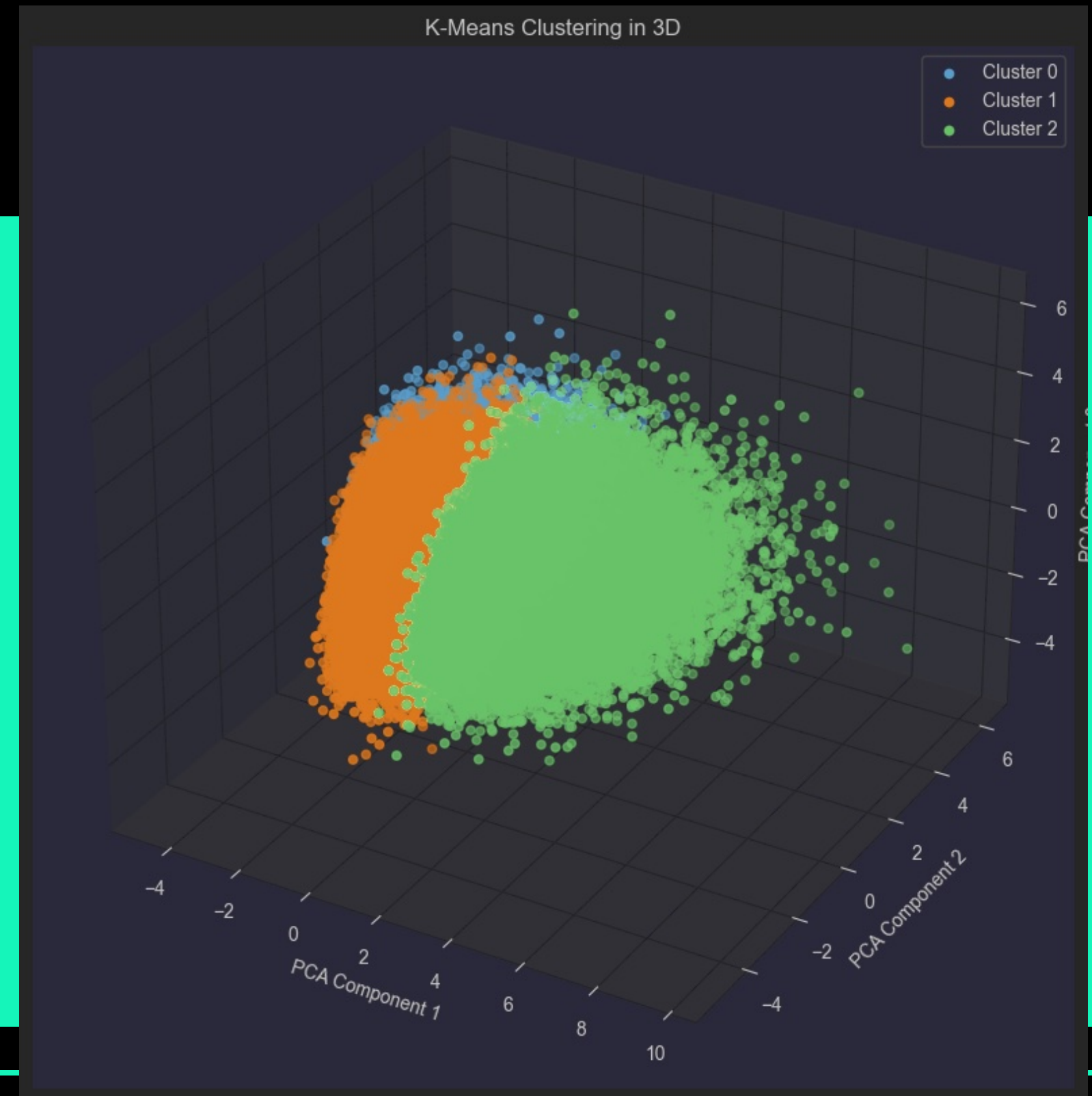
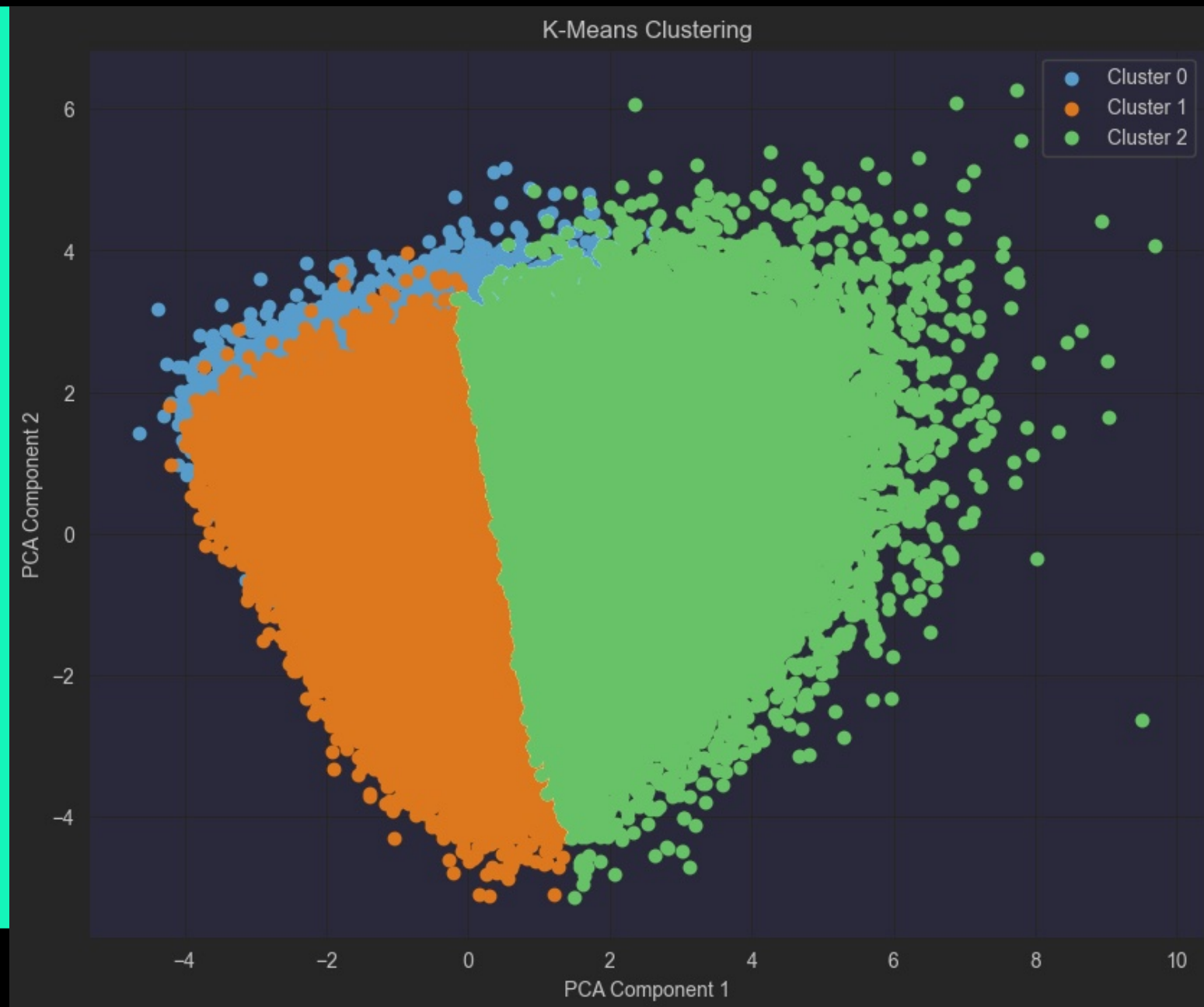
Codo

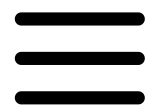


Silueta

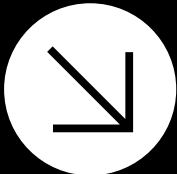


Clustering





Qué variables decidimos mantener



AMT_REQ_CREDIT_BUREAU Cantidad de dudas al Banco por un año	INFO_FLAG Cantidad de datos no enviados	FALSE_INFO Cantidad de información sobre las que el cliente mintió	OCCUPATION_TYPE Tipo de ocupación	CNT_FAM_MEMBERS Cantidad de hijos
DAYS_BIRTH Cumpleaños en días	DAYS_EMPLOYED Días que ha tenido empleo	DAYS_REGISTRATION Días desde que se registró	AMT_INCOME_TOTAL Salario de la persona neto	APP_HOUR_START A qué horas comenzaron la aplicación
NAME_CONTRACT_TYPE Tipo de contrato	FLAG_OWN_CAR Si tiene carro	FLAG_OWN_REALTY Si tiene propiedades	REGION_RATING_CLIENT Califiación de la zona donde habita el cliente	AMT_CREDIT Cantidad del crédito
NAME_INCOME_TYPE Tipo de ingreso	NAME_EDUCATION_TYPE Tipo de educación	NAME_FAMILY_STATUS Estado familiar	TOTALAREA_MODE Moda normalizada del area total de la vivienda del cliente	AMT_ANNUITY Plazos a los que se pide el préstamo
EXT_SOURCE_1 Credit Score 1	EXT_SOURCE_2 Credit Score 2	EXT_SOURCE_3 Credit Score 3		

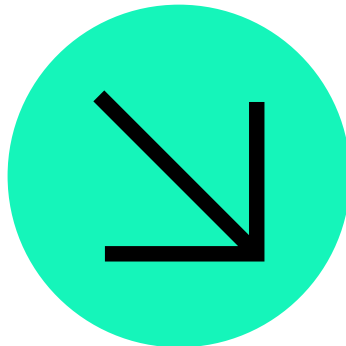
Eliminar atípicos



Eliminar atípicos



Distribución de los datos





Nuestro set de datos se encontraba desbalanceado

La clase minoritaria era menos del

8%

¿Oversampling o Undersampling?

Oversampling:

Es incrementar las instancias de la clase minoritaria por medio de la generación sintética de datos pertenecientes a esta clase

Undersampling:

Reduce las instancias de la clase mayoritaria para igualar la cantidad de datos de la clase minoritaria

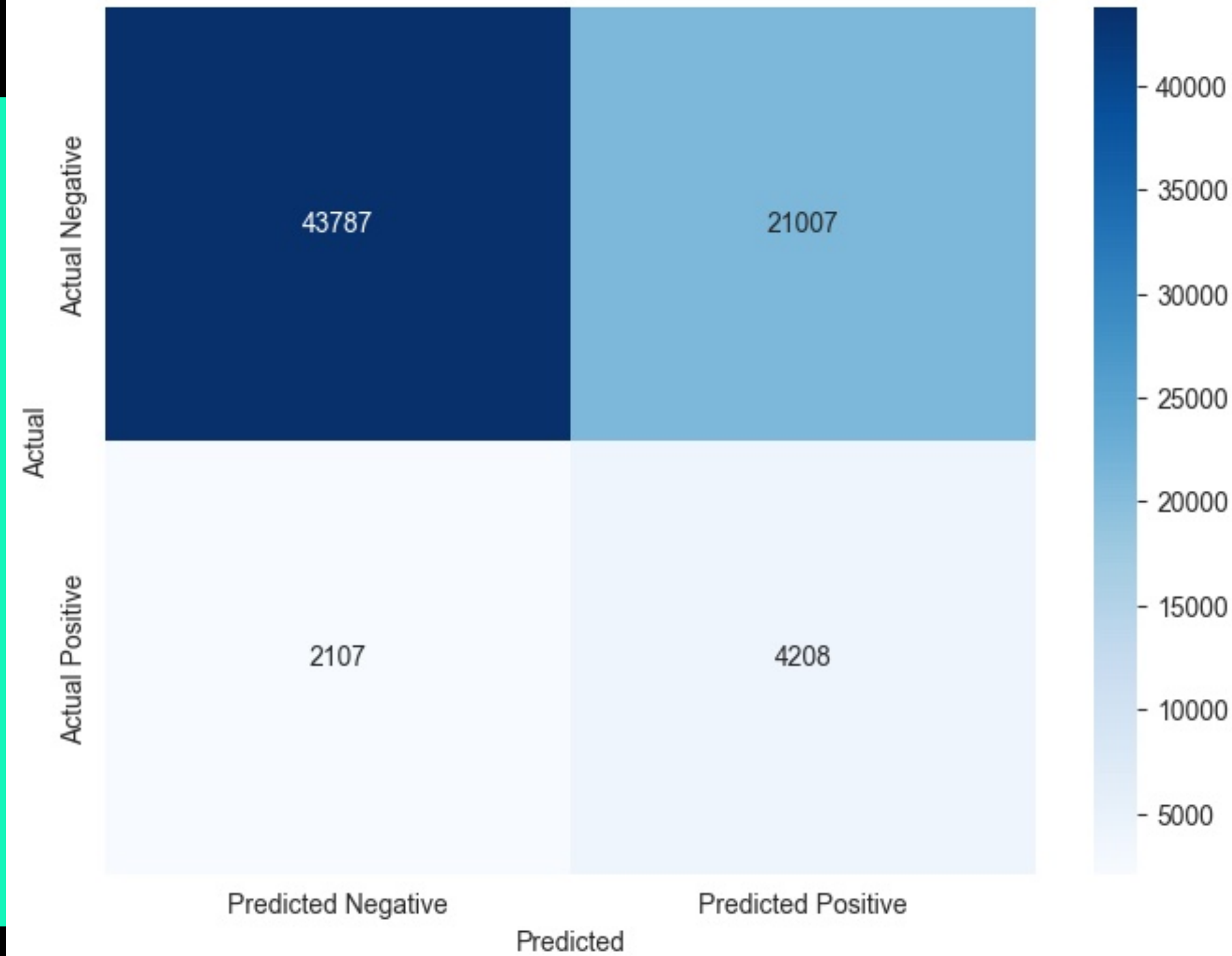
Regresión Logística



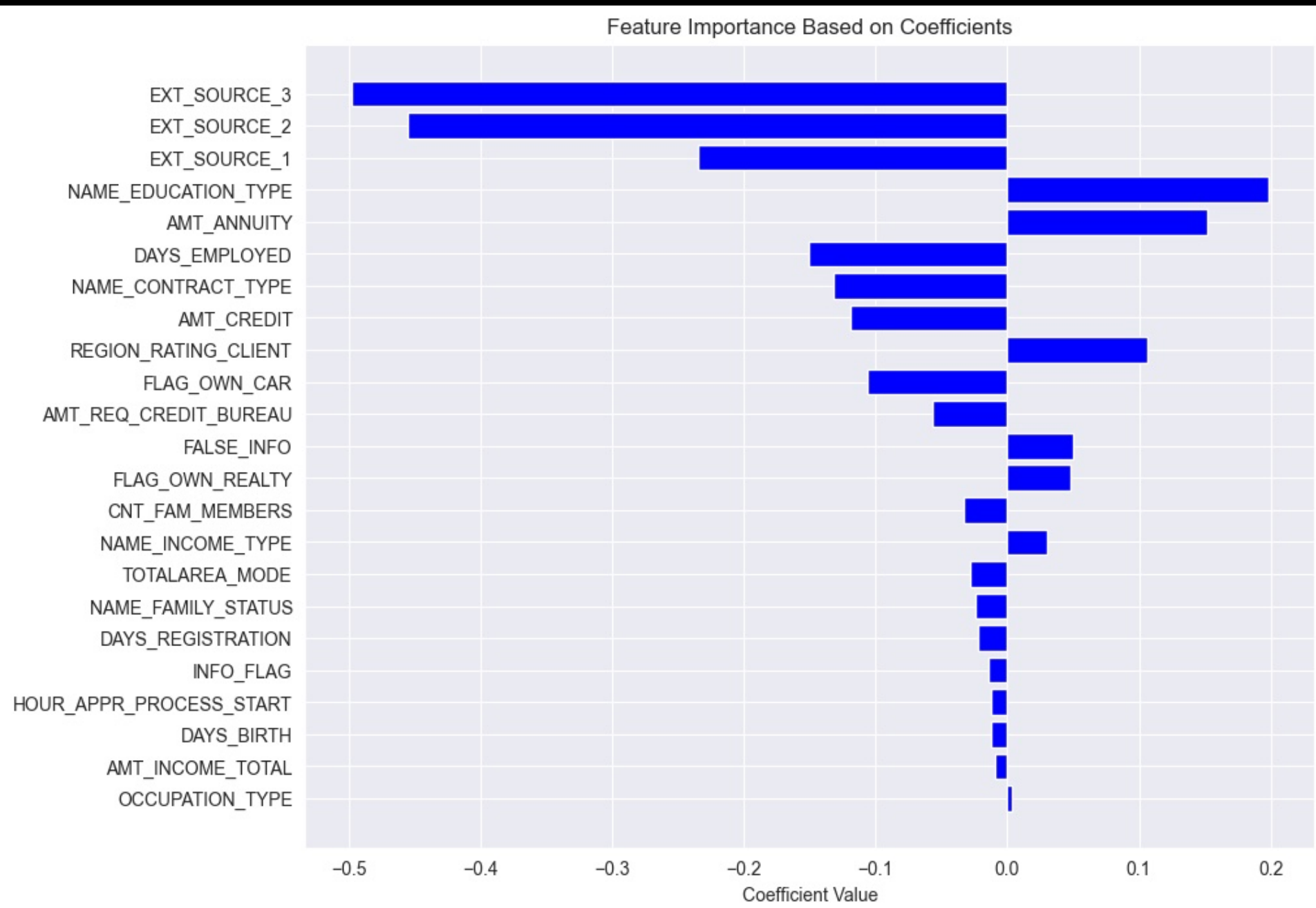
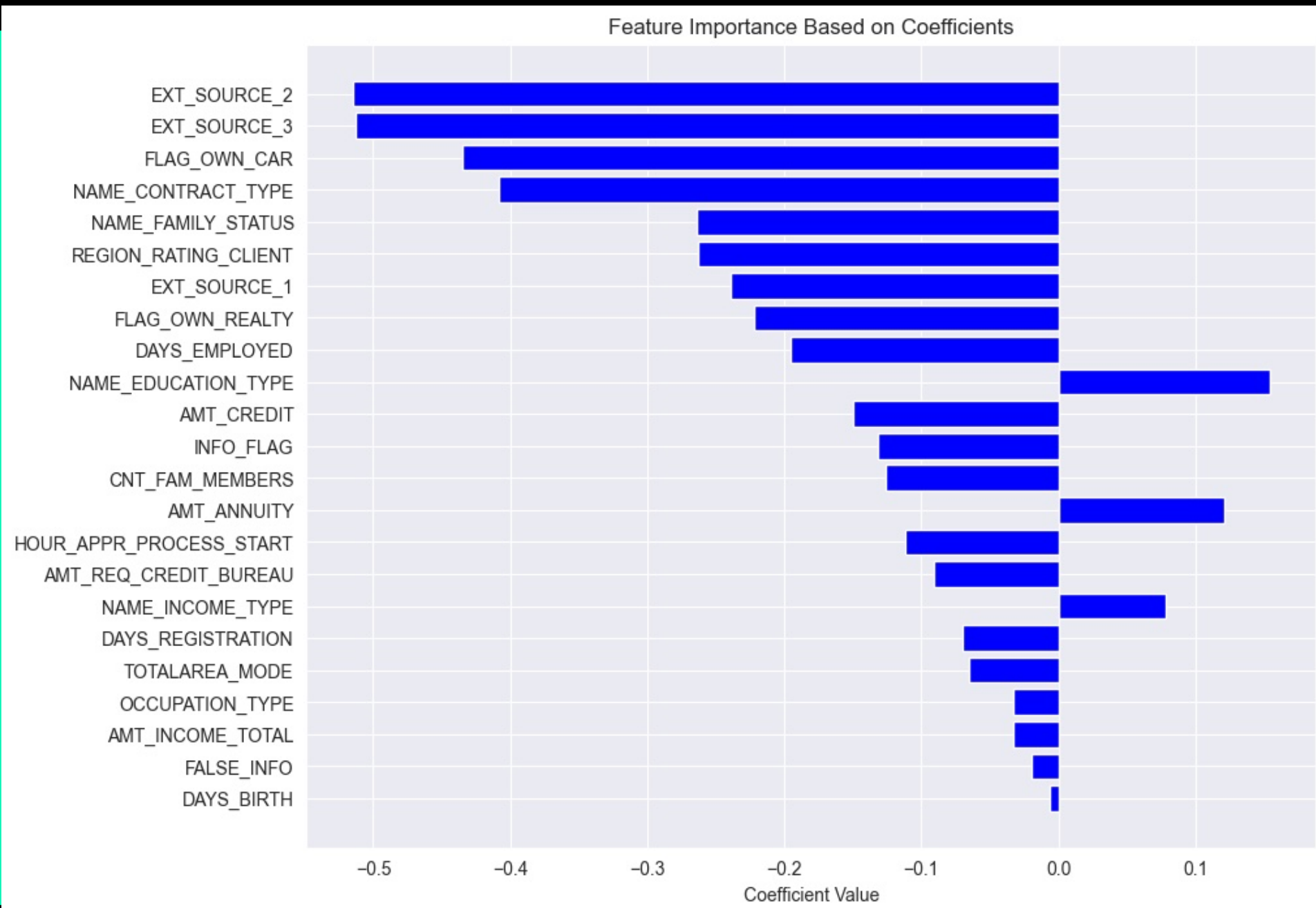
Confusion Matrix



Confusion Matrix



Regresión Logística



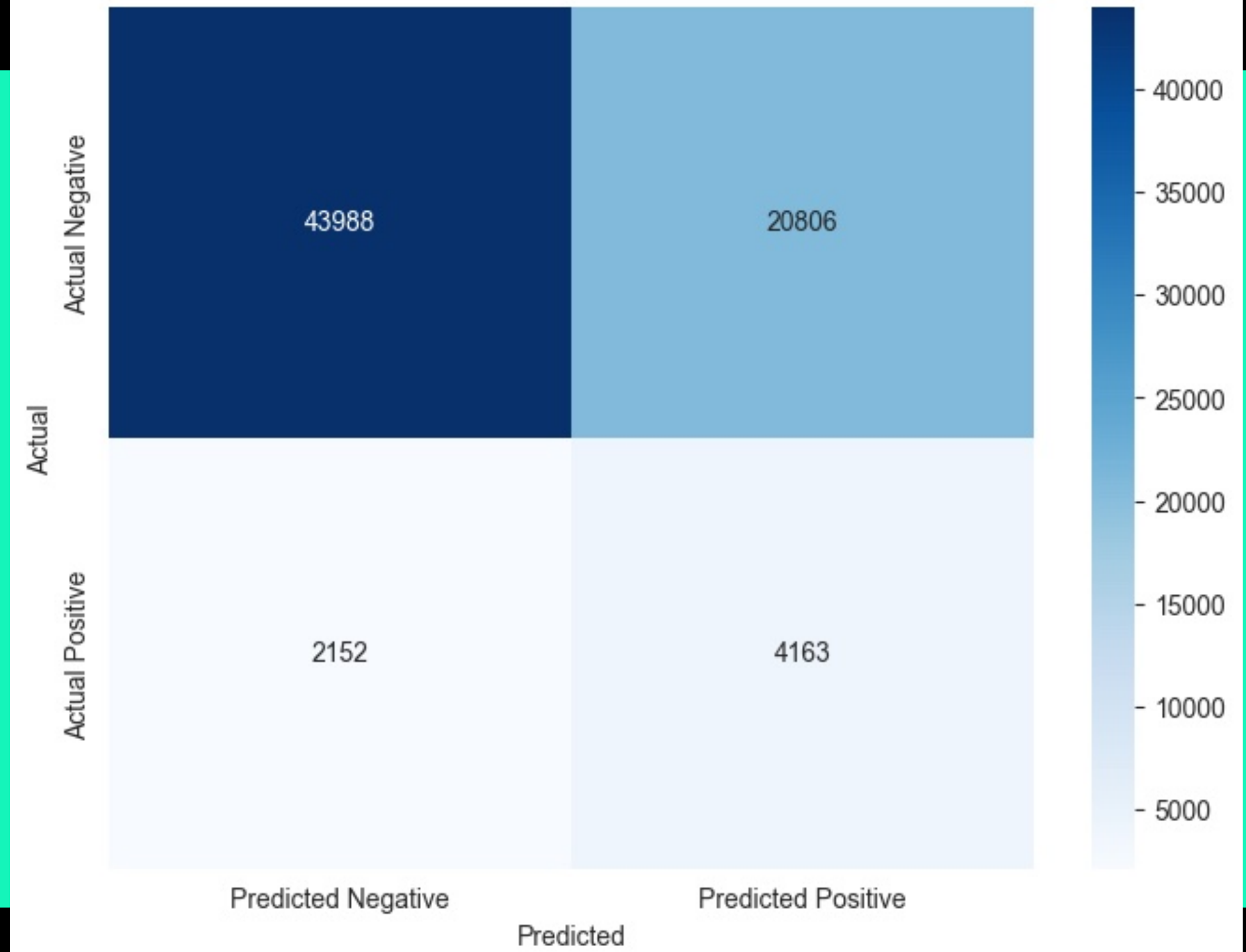
Random forest



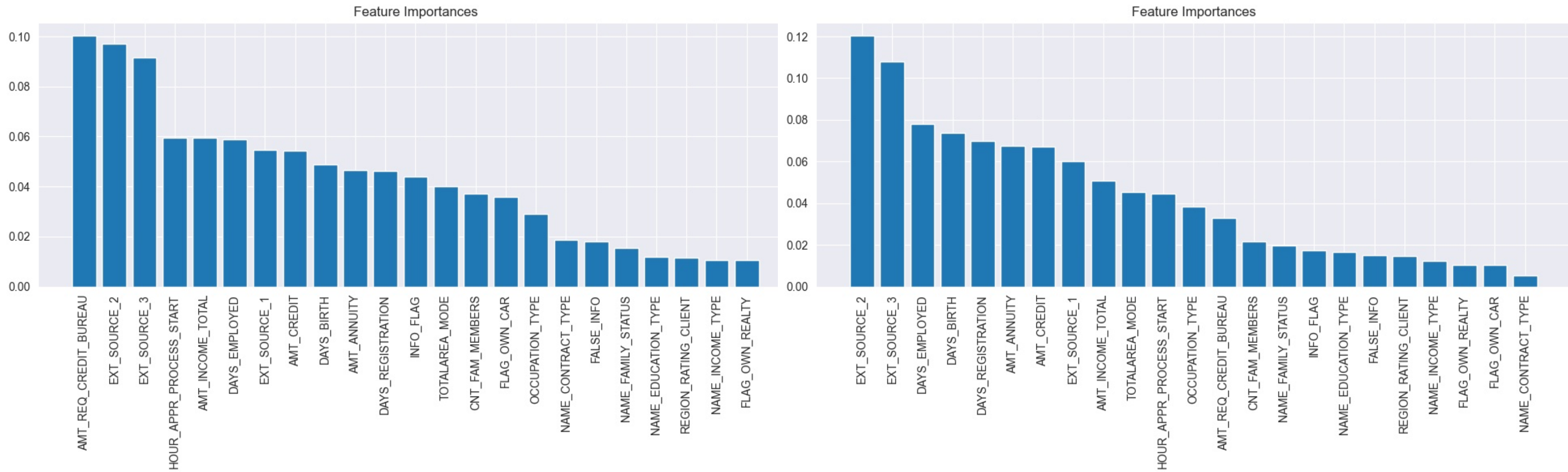
Confusion Matrix



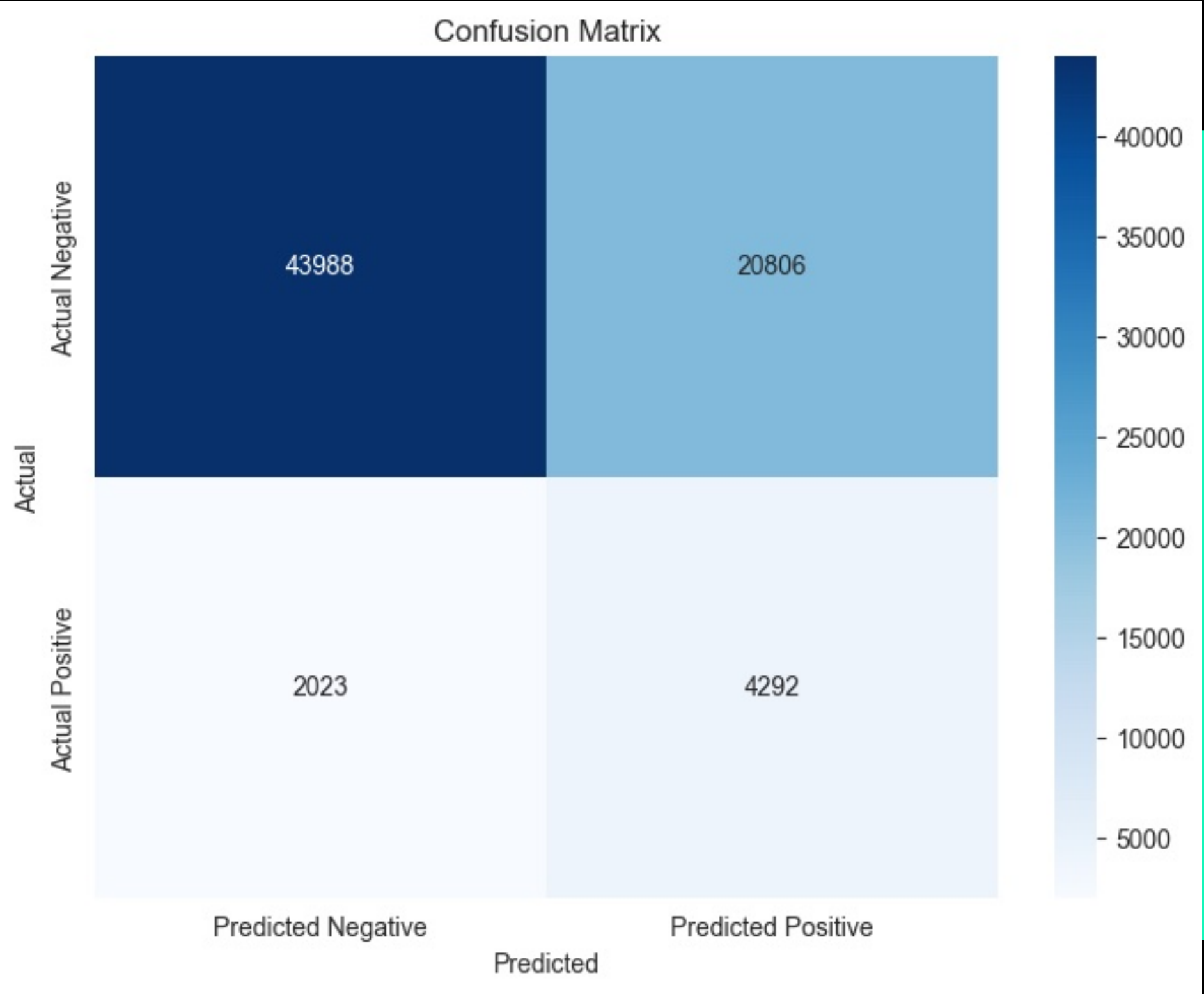
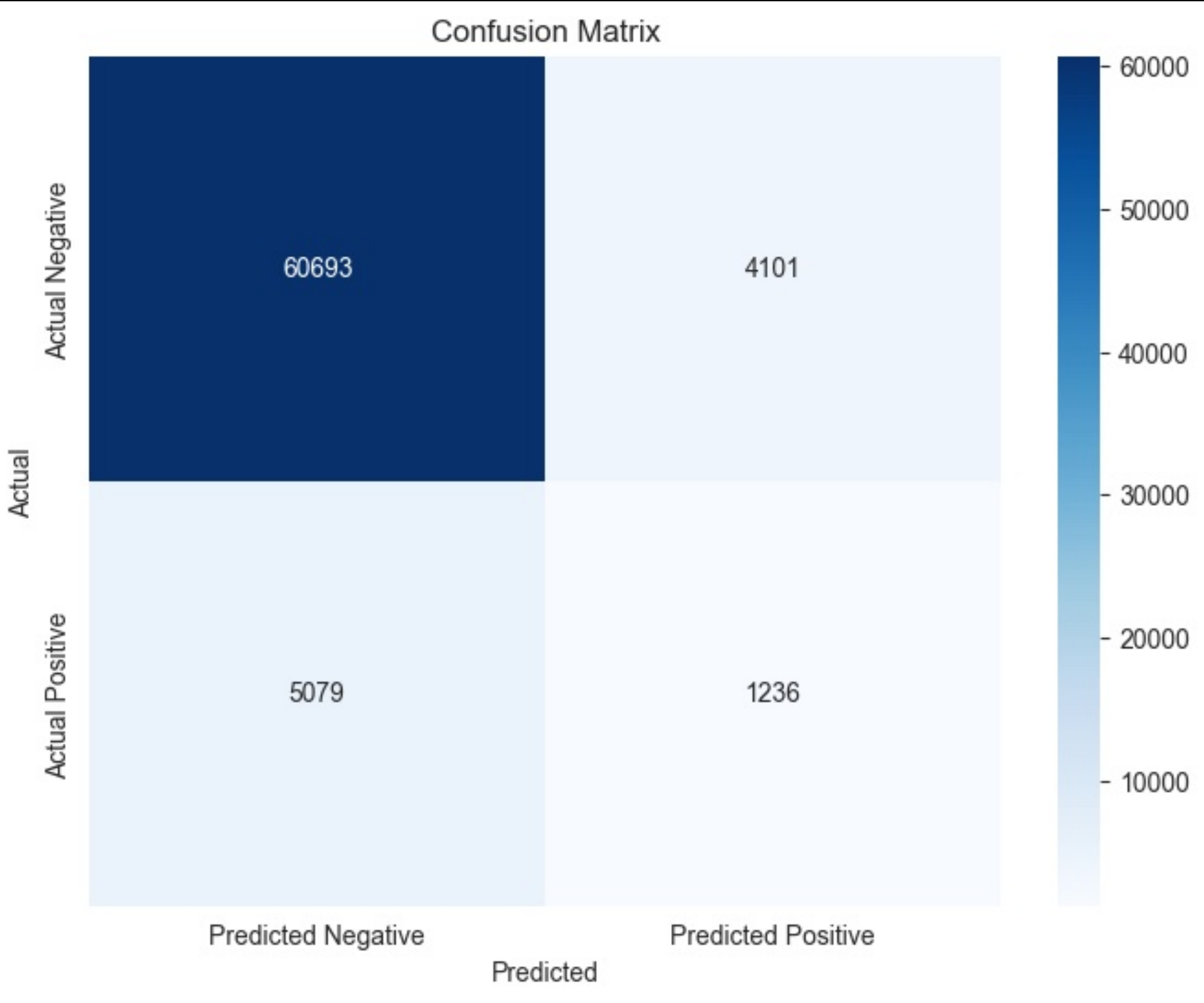
Confusion Matrix



Random forest



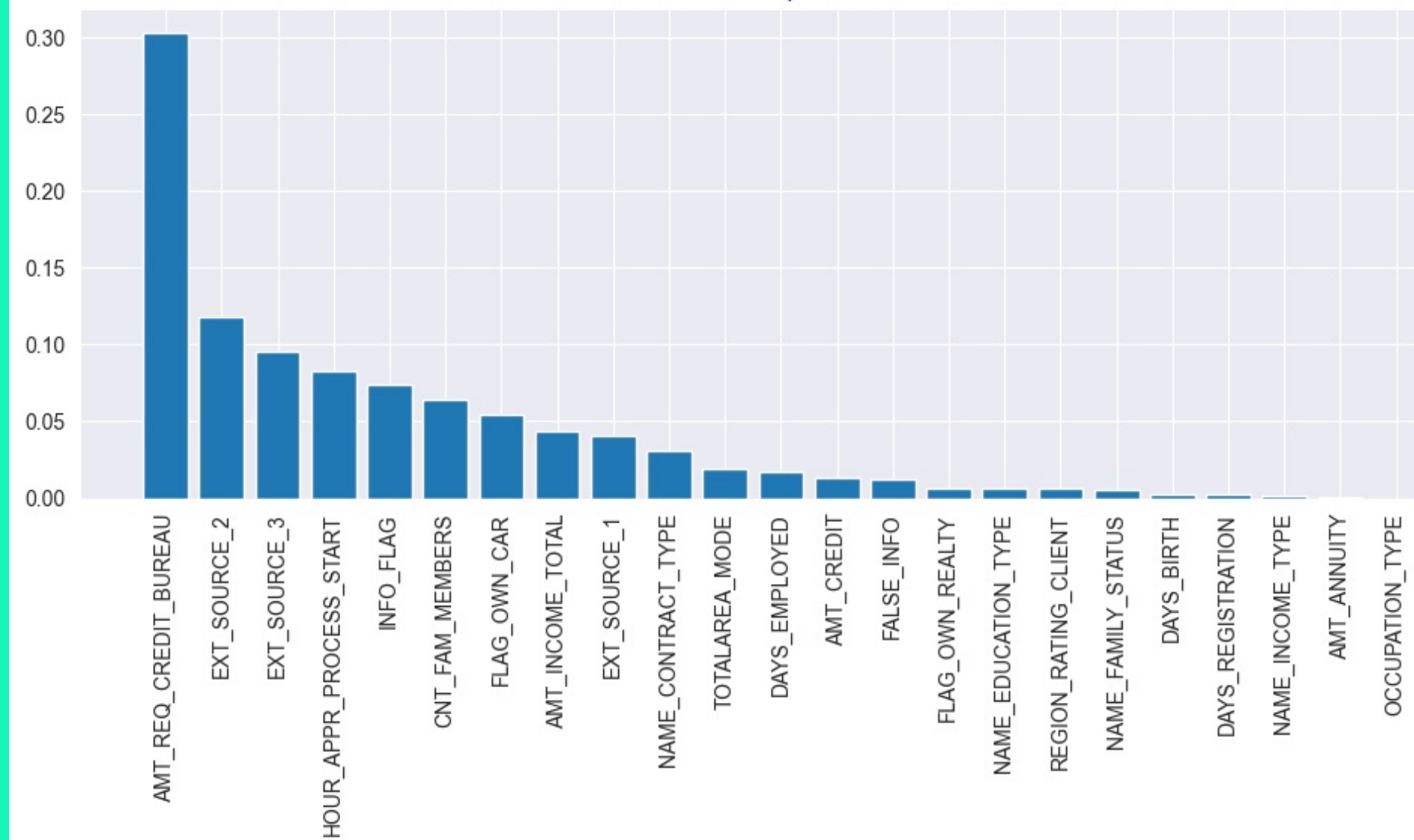
Gradient Boost



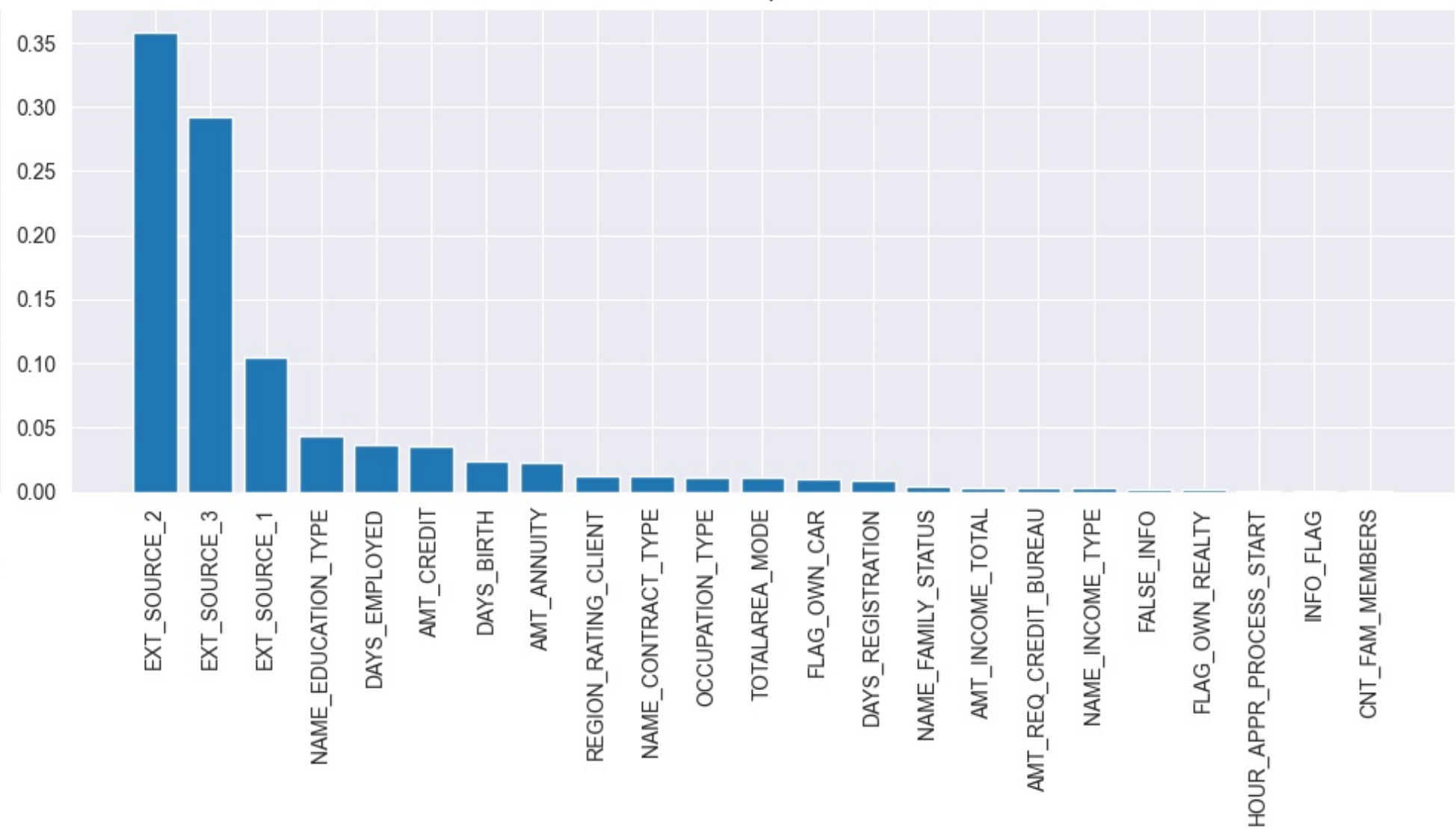
Gradient Boost

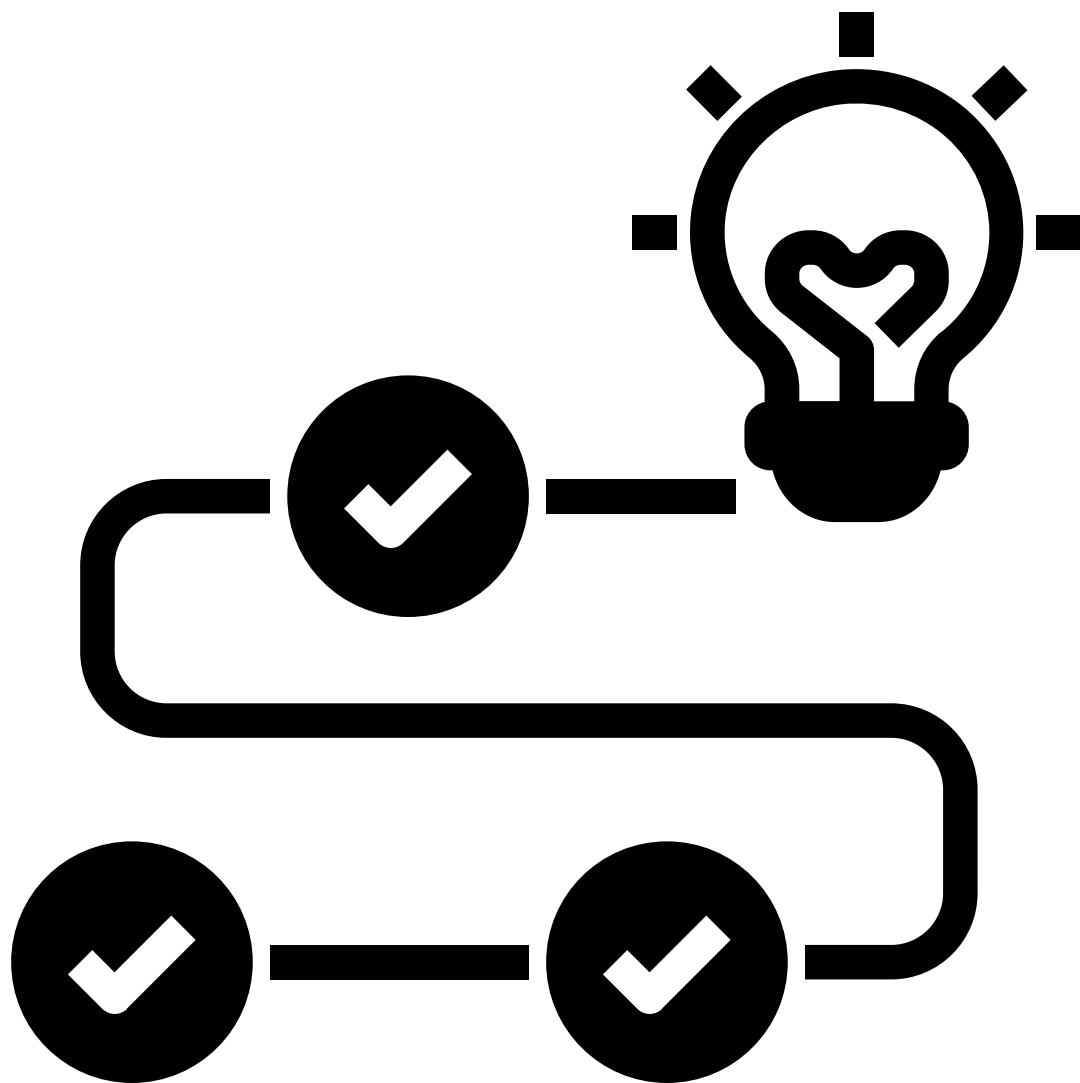


Feature Importances



Feature Importances





Conclusiones

- A veces la técnica adecuada depende de lo que querramos sacrificar
- Un muestreo diferente puede dar resultados muy diferentes
- No logramos hacer un clasificador efectivo



Thank You

