Proyecto Data Science

Grupo 5: Desafío Banca

2022

Hito 3

1.1 - Presentación del equipo de trabajo:

Integrantes:	**Roles:**
María Teresa Moncada Moraga	Líder, Analista de datos.
Felipe Riveros Fuentes	Ingeniero de modelamiento, Visualización de datos.
José Luis Pozo Saavedra	Ingeniero de modelamiento, Visualización de datos.
Nicolas Vasquez Mujica	Ingeniero de modelamiento.

Todos los integrantes aportan en los roles de documentador, control de calidad y validación de datos.

1.2 - Descripción del tema y motivación:

Se elige trabajar con un tema propuesto por la academia el cual se llama "Otorgar nuevos créditos".
Este tema consiste en una entidad financiera (Banco Internacional) la cual otorga créditos tanto de consumo como revolving y encomienda al área de Data Science encontrar un modelo para **predecir el comportamiento de futuros buenos pagadores**.

Descripción: El tema escogido, consiste principalmente en un problema de clasificación, debido a que la variable objetivo variable cualitativa y de carácter nominal, para lograr predecir si una persona podría ser un buen pagador y calificar para un nuevo crédito. Dentro del problema nos da a escoger entre la cartera de consumo o revolving, escogeremos consumo, porque tiene mayor cantidad de datos.

Motivación: Es un tema interesante y muy presente en la industria financiera, como usuario sería relevante entender cuales son las posibles variables que se tienen en cuenta al momento de evaluar un crédito y que posibilidades tendría el usuario de acceder a un crédito de acuerdo a ciertas condiciones. Por el lado del Banco es una oportunidad de hacer más eficiente el proceso de evaluación crediticia, tanto para clientes nuevos como para clientes antiguos del Banco. Y por último, consideramos este proyecto interesante al estar estrechamente ligado a trabajos posibles para Cientistas de Datos, acercándonos de alguna forma al mundo laboral real.

1.3 - Planificación de la investigación:

Hipótesis de investigación: Las variables económicas, son más determinantes que las variables sociales para optar a un nuevo crédito.

Vector Objetivo: Target.

Objetivo General: Predecir si la persona es apta para un nuevo crédito.

Objetivos Específicos:

- Identificar correlaciones entre las variables que podrían inferir en el resultado.
- Entrenar modelos predictivos que permitan estimar la probabilidad de buen comportamiento de los clientes.

Pregunta de investigación: ¿Cuáles son las variables que influyen en un buen comportamiento de pago?

Estrategia: Para este problema se realizará una limpieza de los datos, se inspeccionarán los nulos y se reemplazarán para perder la menor cantidad de datos. Las variables categóricas serán binarizadas. La variable objetivo "TARGET", puede tomar sólo dos valores 0 y 1 (Buen pagador y mal pagador respectivamente), por lo que es un tipo de variable cualitativa y de carácter nominal. Debido a lo anterior, para resolver este problema se utilizará un modelo de clasificación validado desde la econometría y un modelo predictivo que será validado desde Machine Learning.

Encriptación de datos sensibles. ¿Qué datos serán transformados?

No se visualizan datos sensibles en el dataset.
Las variables de prefijo “FLAG_” serán recodificadas con 0 y 1.

Tratamiento de datos faltantes: ¿Qué técnicas de imputación se usarán o se eliminarán los registros?

Se quitarán los registros de Revolving, filtrando por el campo “NAME_CONTRACT_TYPE” = “Cash loans”.

Muestreo: ¿Se trabajará con la población o una muestra? Explicar el motivo de la decisión.

Se trabajará con el dataset de modelación entregado por la academia la cual consiste en dos archivos csv. El primero Desafio4_modelamiento.csv y el segundo archivo es Desafio4_Validacion.csv.
Inicialmente se plantea trabajar con una submuestra balanceada la cual se aplicaría solo a la muestra de modelamiento, esto se cree porque tenemos un 8% de casos de Target = 1.

Análisis exploratorio: ¿Existen patrones importantes en los datos?, ¿Qué tipos de datos se tienen?, ¿Qué tipo de dato es el vector objetivo?, ¿Dispersión y distribución de los datos?, entre otros.

- Se detecta un desbalance de la clase Target.
- El vector objetivo es una variable cualitativa y de carácter nominal.
- Modelación adecuada para el vector objetivo es de tipo clasificación.

Dentro de este apartado se realizará todo el análisis descriptivo y de características del conjunto de datos, usando las herramientas de comprensión de antecedentes como Medidas de posición central, medidas de dispersión, correlaciones y gráficos.

Propuesta de modelos a utilizar:

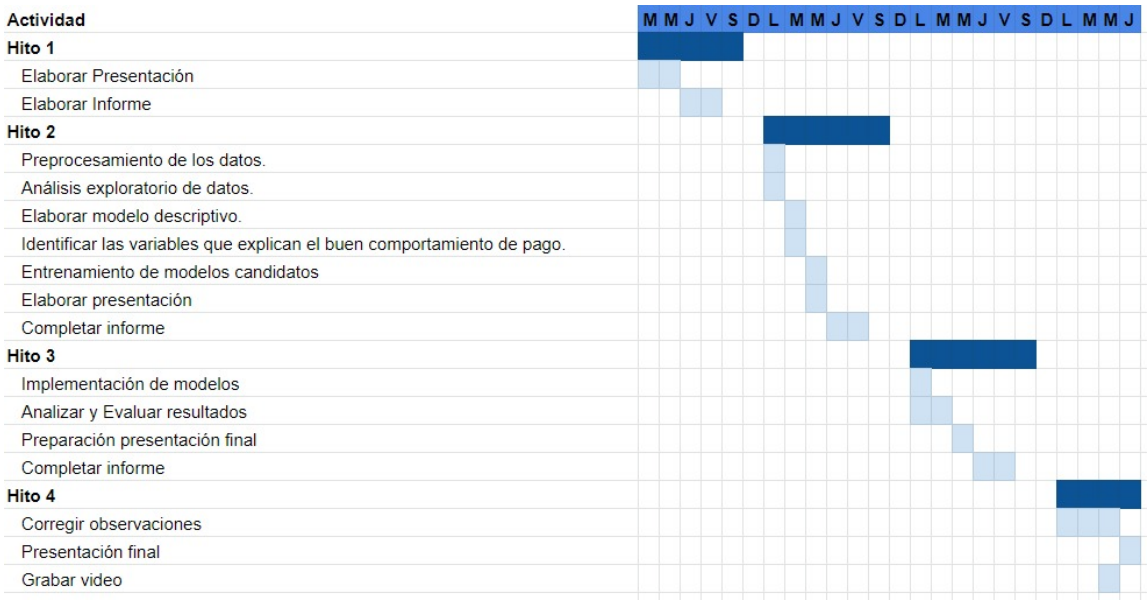
1. Regresión Logística
2. Random Forest
3. Adaboost, Gradientboost
4. Naive Bayes
5. SVM

Planificación Proyecto

Para realizar este proyecto de data science se consideran los siguientes hitos de planificación.

- Preprocesamiento de los datos.
- Análisis exploratorio de datos.
- Elaboración de modelos descriptivos.
- Entrenamiento de modelos candidatos.

A continuación se muestra una planificación del proyecto, en donde se indican los hitos entregables para cada semana y su estimación de duración.



2.1 - Análisis descriptivo y gráficos relevantes

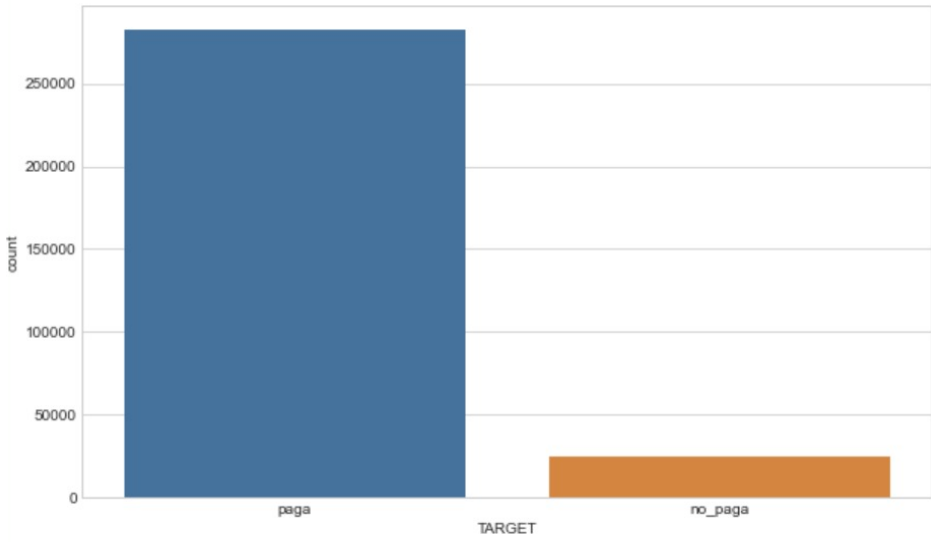
En primera instancia se revisan los archivos de datos entregados por la academia, los cuales se llaman Desafio4_modelamiento.csv y Desafio4_Validacion.csv. los cuales tienen un total de 122 variables que hemos segregado en los siguientes tipos:

- Variables Objetivo: Target
- Variables ID: 2
- Variables Bancarias: 37
- Variables sociodemográficas: 82

Se realizó un análisis descriptivo de las variables más importantes y del vector a predecir, con el objetivo de entender a priori su comportamiento en términos de dispersión, relación entre variables predictoras y describir posibles tendencias.

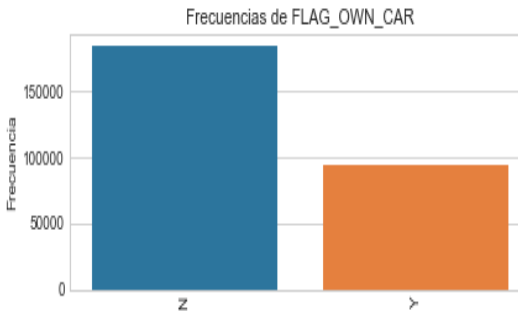
Análisis de variables relevantes

1- Gráfico variable objetivo

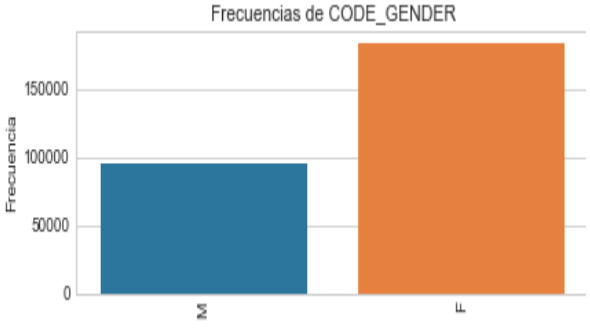


Se observa que la cantidad de personas con dificultades para pagar son mucho menos que las personas que no tienen dificultades, tan sólo 23.221 lo que representa un 8% del total.

2- Gráfico Vehiculo



3 - Sexo cliente



De los gráficos 2 y 3 podemos observar que la cantidad de personas que tenemos en el archivo de datos es mayoritariamente femenino y sin automóvil.

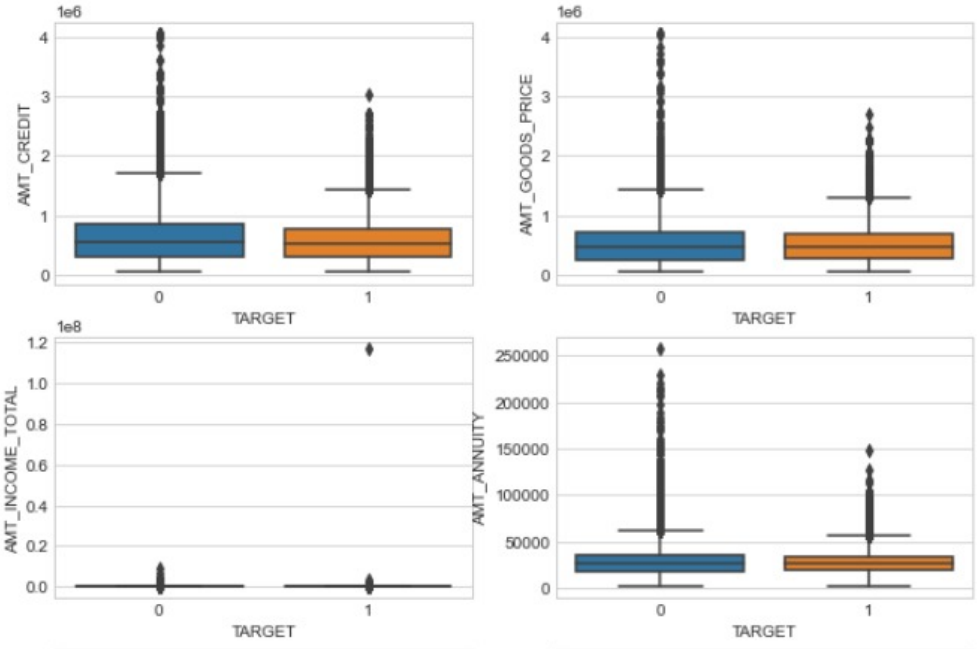
2.2 - Limpieza, estructuración de datos y Feature

Una vez realizado el estudio de las variables en el primer archivo de datos encontramos que en el archivo de validación no se adjuntaba la variable objetivo, y posterior a conversación con él profesor se nos aconsejo trabajar solo con el primer archivo y dividir a este en muestras tanta para entrenamiento (70%) como para validación (30%).

Una vez definido el universo y las respectivas muestras, se procede a realizar un análisis de datos faltantes en las variables para proceder si amerita la eliminación de las mismas.

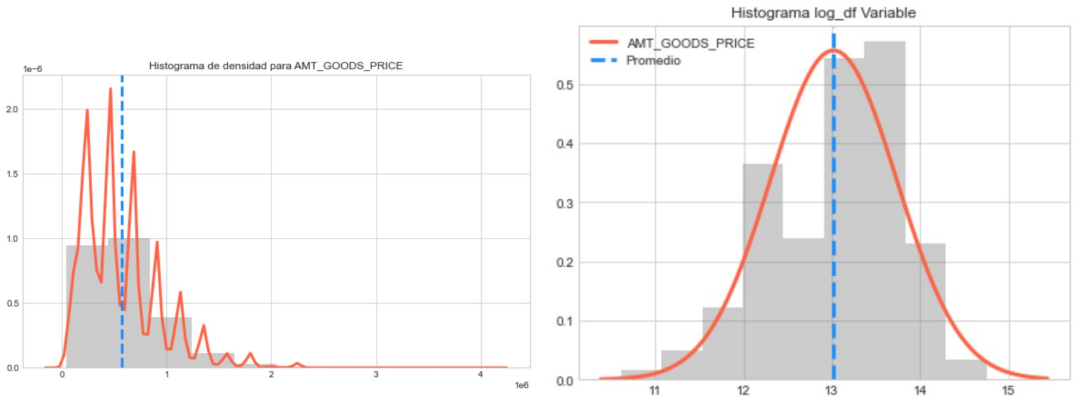
Se realiza un análisis de datos perdidos, encontrando un total de 49 variables candidatas a ser eliminadas por tener sobre un 48% de missing. En anexo se encuentra el listado completo de variables, su % de missing. Además se eliminan 36 variables por corresponder a identificadores (ID), por no tener una descripción clara de lo que significan o por ser irrelevantes al estudio. En anexo se encuentra la explicación de eliminación de cada una de las variables.
Se realiza el análisis de datos outlier encontrando las siguientes observaciones:

5. - Gráfico outlier



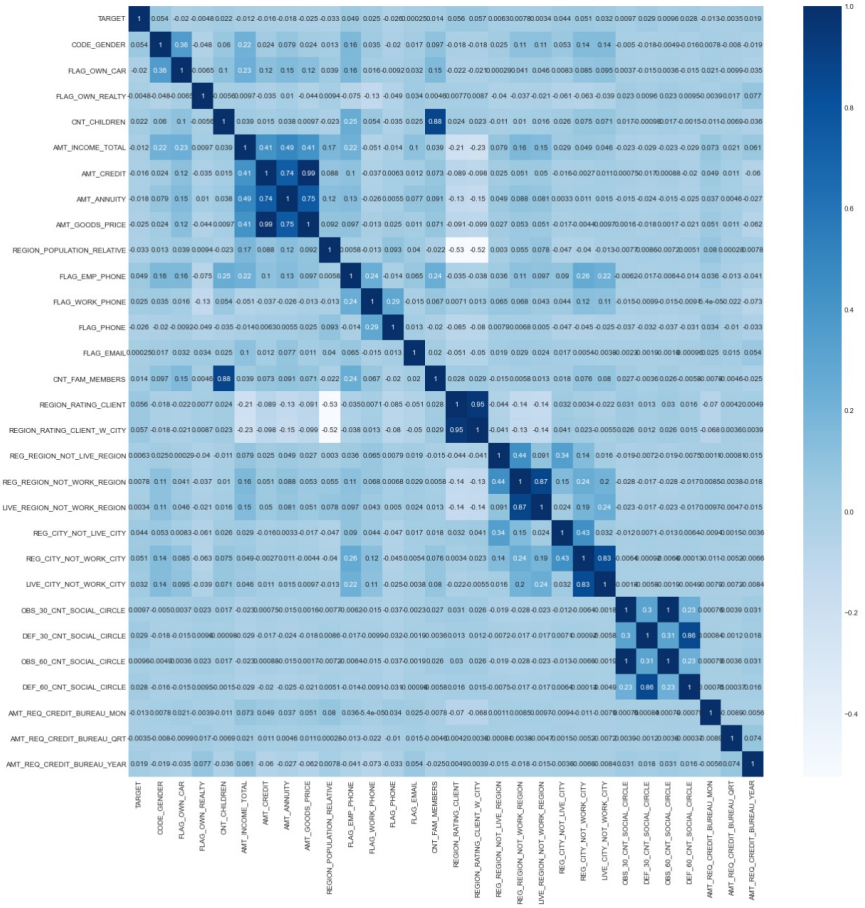
En el gráfico de outlier, se observa que la variable days_employed tiene valores negativos y valores muy altos que representa haber trabajado sobre 1000 años (>300.000 días), lo cual no podría ser considerado un dato válido. Por otra parte, la edad del cliente (days_birth) mantiene sólo valores negativos.

6. - Gráfico AMT_GOODS_PRICE



Se estudian gráficamente las variables continuas principalmente asociadas a montos, donde podemos observar a modo de ejemplo, en el primer gráfico que el “precio de los bienes que se comparan con el préstamos” (AMT_GOODS_PRICE) la variable tiene una tendencia marcada asintóticamente a la izquierda, y en el segundo gráfico se observa la misma variable normalizada (logaritmo)

7. - Gráfico Correlaciones



Se realiza un análisis de correlaciones entre las variables objetivo(Target) y el resto de las variables en el dataset, no encontrando ninguna variable altamente correlacionada con el vector objetivo.

Después de los análisis realizados a los archivos de datos podemos decir que existen variables que serán recodificadas, normalizadas y otras eliminadas.

Se recodifican las siguientes variables:

- OCCUPATION_TYPE: Se recodifican los valores perdidos en una nueva categoría definida como “desconocido” (UNKNOWN)
- Variables binarias con respuesta (Y or N) por (1 o 0)
- Variables binarias con respuesta (M o F) por (1 o 0)
- Variables categóricas, son binarizadas.
- Variables como ingreso (AMT_INCOME_TOTAL) son normalizadas (logaritmo).

Bases de datos y variables utilizadas

Se utilizaron un total de 122 variables, de las cuales por las distintas técnicas analizadas se eliminaron 87 variables, quedando un dataset final de 35 variables candidatas (atributos) para predecir nuestro target. A continuación, que se presentan las variables que serán utilizadas en los diferentes modelos predictivos:

● **TARGET**: Variable objetivo (1 - El cliente presenta dificultades para pagar: Presenta un pago atrasado más de X días en por lo menos en los primeros meses de pago).

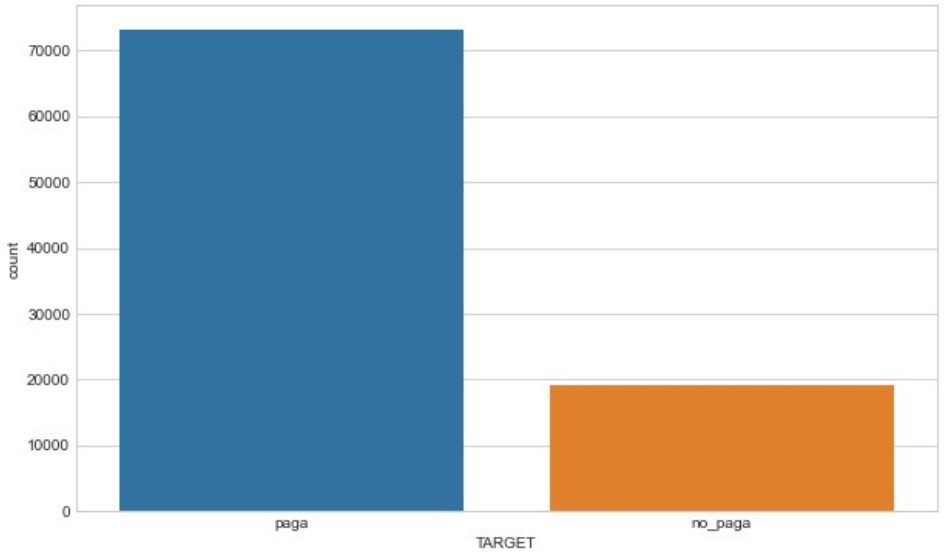
● **CODE_GENDER**: Sexo del cliente.
● **FLAG_OWN_CAR**: Indicador binario sobre la tenencia de automóvil por parte del cliente.
● **FLAG_OWN_REALTY**: Indicador binario sobre la propiedad de una casa o departamento por parte del cliente.
● **CNT_CHILDREN**: Cantidad de hijos por parte del cliente.
● **AMT_INCOME_TOTAL**: Ingreso total del cliente.
● **AMT_CREDIT**: Cantidad total del préstamo realizado.
● **AMT_ANNUITY**: Anualidad del préstamo.
● **AMT_GOODS_PRICE**: Para préstamos de consumo representa el precio de los bienes que se comprara con el préstamo.
● **NAME_INCOME_TYPE**: Tipo de ingreso por parte del cliente (empresario, asalariado, postnatal, entre otros).
● **NAME_EDUCATION_TYPE**: Máximo nivel educacional por parte del cliente.
● **NAME_FAMILY_STATUS**: Situación familiar del cliente.
● **NAME_HOUSING_TYPE**: Cuál es la situación habitacional del cliente. (arrienda, allegado, entre otros).
● **REGION_POPULATION_RELATIVE**: Población donde vive el cliente (la variable está normalizada, donde valores más altos significan que el cliente vive en una región más poblada).
● **FLAG_EMP_PHONE**: Da un teléfono de trabajo de contacto:○ 1: YES.○ 0: NO.
● **FLAG_WORK_PHONE**: Da un teléfono de hogar de contacto:○ 1: YES.○ 0: NO.
● **FLAG_PHONE**: Da un teléfono contacto el cliente: ○ 1: YES.○ 0: NO.
● **FLAG_EMAI**: Da un email de contacto el cliente:○ 1: YES○ 0: NO
● **OCCUPATION_TYPE**: Cuál es la profesión del cliente.

● **CNT_FAM_MEMBERS**: Cuántos miembros familiares tiene el cliente.
● **REGION_RATING_CLIENT**: Evaluación interna (de Home Credit Group) sobre la región donde vive el cliente.
● **REGION_RATING_CLIENT_W_CITY**: Evaluación interna (de Home Crédito Group) sobre la región donde vive el cliente considerando ciudad.
● **REG_REGION_NOT_LIVE_REGION**: Identificador booleano si es que la dirección permanente del cliente no concuerda con la dirección de contacto:
○ 1: different.○ 0: same, at region level.
● **REG_REGION_NOT_WORK_REGION**: Identificador booleano si es que la dirección permanente del cliente no concuerda con la dirección de trabajo:
○ 1: different.○ 0: same, at region level).
● **LIVE_REGION_NOT_WORK_REGION**: Identificador booleano si es que la dirección de contacto del cliente no concuerda con la dirección del trabajo:
○ 1: different.○ 0: same, at region level).
● **REG_CITY_NOT_LIVE_CITY**: Identificador booleano si es que la dirección permanente no concuerda con la dirección de contacto:
○ 1: different.○ 0: same, at city level.
● **REG_CITY_NOT_WORK_CITY**: Identificador booleano si es que la dirección permanente no concuerda con la dirección del trabajo:
○ 1: different.○ 0: same, at city level.
● **LIVE_CITY_NOT_WORK_CITY**: Identificador booleano si es que la dirección de contacto del cliente no concuerda con la dirección del trabajo:
○ 1: different.○ 0: same, at city level.
● **ORGANIZATION_TYPE**: Tipo de organización donde trabaja el cliente.
● **OBS_30_CNT_SOCIAL_CIRCLE**: Cuántas veces ha registrado mora más de 30 días su entorno.
● **DEF_30_CNT_SOCIAL_CIRCLE**: Cuántas veces ha registrado mora más de 30 días su entorno.
● **OBS_60_CNT_SOCIAL_CIRCLE**: Cuántas veces ha registrado mora más de 60 días su entorno.
● **DEF_60_CNT_SOCIAL_CIRCLE**: Cuántas veces ha registrado mora más de 60 días su entorno.
● **AMT_REQ_CREDIT_BUREAU_MON**: Cantidad de consultas sobre el cliente al buró de crédito. Un mes antes de la postulación.
● **AMT_REQ_CREDIT_BUREAU_QRT**: Cantidad de consultas sobre el cliente al buró de crédito. Tres meses antes de la postulación.
● **AMT_REQ_CREDIT_BUREAU_YEAR**: Cantidad de consultas sobre el cliente al buró de crédito. Un año antes de la postulación.

Balanceo de la muestra

Para resolver la problemática del balanceo, se revisaron distintas opciones, resolviendo que lo más adecuado era generar una submuestra aleatoria de la clase mayoritaria, disminuyendo la cantidad de casos y manteniendo los registros de la clase minoritaria. Como resultado, el dataset pasó de tener cerca de 240.000 registros a 100.000 aproximadamente y el desbalance pasó de 92/8 a un 80/20.

8. Gráfico Target



2.3 - Entrenamiento de modelos candidatos.

Una vez obtenida la muestra, se utiliza un seudo balanceo reduciendo la clase mayoritaria y obteniendo una submuestra aleatoria. Con esta muestra se ejecutarán los 6 modelos antes mencionados: Regresión Logística, Random Forest, Adaboost, Gradientboost, Naive Bayes y SVM. Estos modelos se eligen por ser los modelos que más trabajamos durante la carrera.
Los hiperparametros que se usarán son los siguientes:

Modelos	Hiperparámetro
Regresión Logística	C:[0,0001 0.001, 0.01, 0.1] penalty:['l1', 'l2', 'none'] 'class_weight': ['balanced'] 'solver':['newton-cg']
Naive Bayes	alpha = [0.1, 0.5, 1] fit_prior=[True,False]
Random Forest	n_estimators = [50, 200, 500] max_features = [0.5, 0.7, 'log2'] max_depth = [3,6, 8], class_weight = ['balanced']

	bootstrap: [False, True] criterion: ['entropy']
SVM	C = [0.0001, 0.001, 0.01, 0,1, 0.5, 1, 5, 10] gamma = [0.0001, 0.001, 0.01, 0,1, 0.5, 1, 5, 10]
Gradient Boosting	n_estimators = [100, 500, 1000] learning_rate = [0.01, 0.1, 0.5, 1, 5, 10] subsample = (0.1, 1.0, 0.5)
Adaptative Boosting	n_estimators = [100, 1000, 10] learning_rate = [0.01, 100, 10]

En primera instancia se realizan los modelos sin hiperparametros y se comparara en posterior ejecución los mismos modelos con la grilla de parámetros presentada anteriormente.

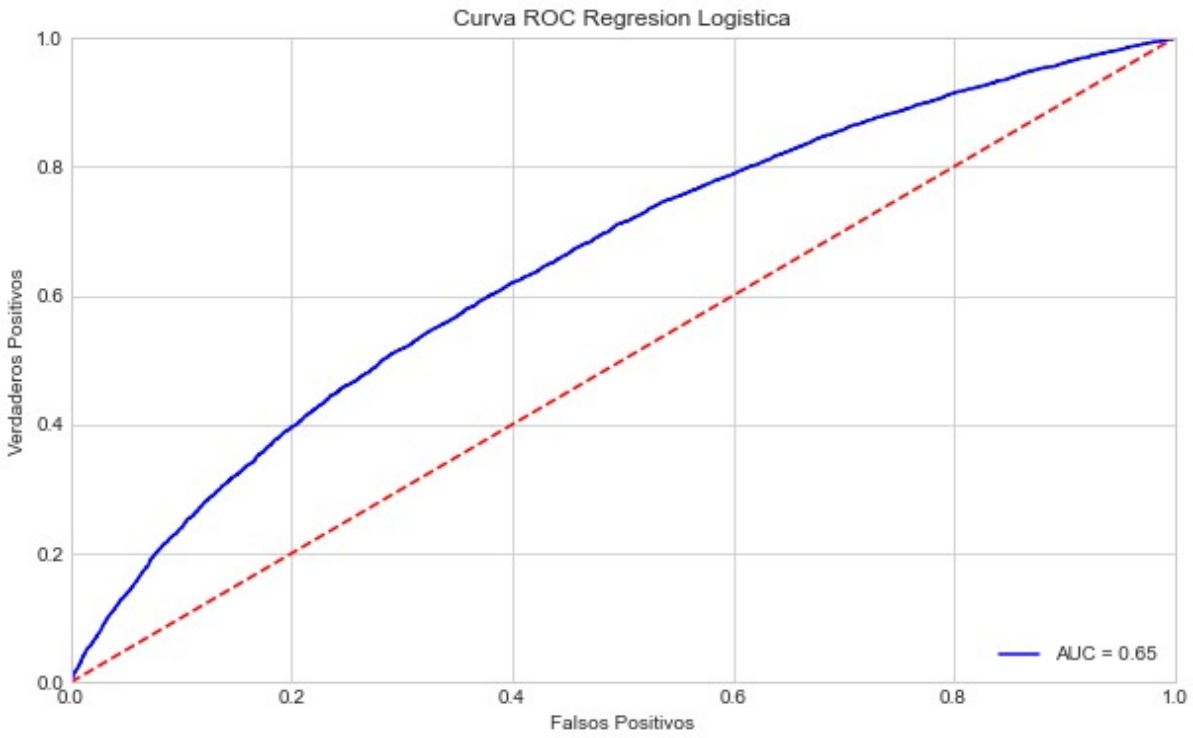
Resultado modelos con hiperparametros y muestra pseudobalanceada

2.3.1 - Regresión Logística

	precision	recall	f1-score	support
0	0.86	0.62	0.72	21976
1	0.29	0.60	0.39	5748
accuracy			0.61	27724
macro avg	0.57	0.61	0.55	27724
weighted avg	0.74	0.61	0.65	27724

El Modelo Regresión Logística con hiperparametros consigue una precisión de 74%.
Para el f1, en los casos 0 es de 0.71 y para los casos en que el cliente no paga el crédito(1) es de 0.39.

Si bien la precisión del modelo es menor a la del modelo sin parámetros en este escenario podemos predecir de mejor manera los casos en que caen en incumplimiento los clientes del banco en comparación al mismo modelo sin parámetros.

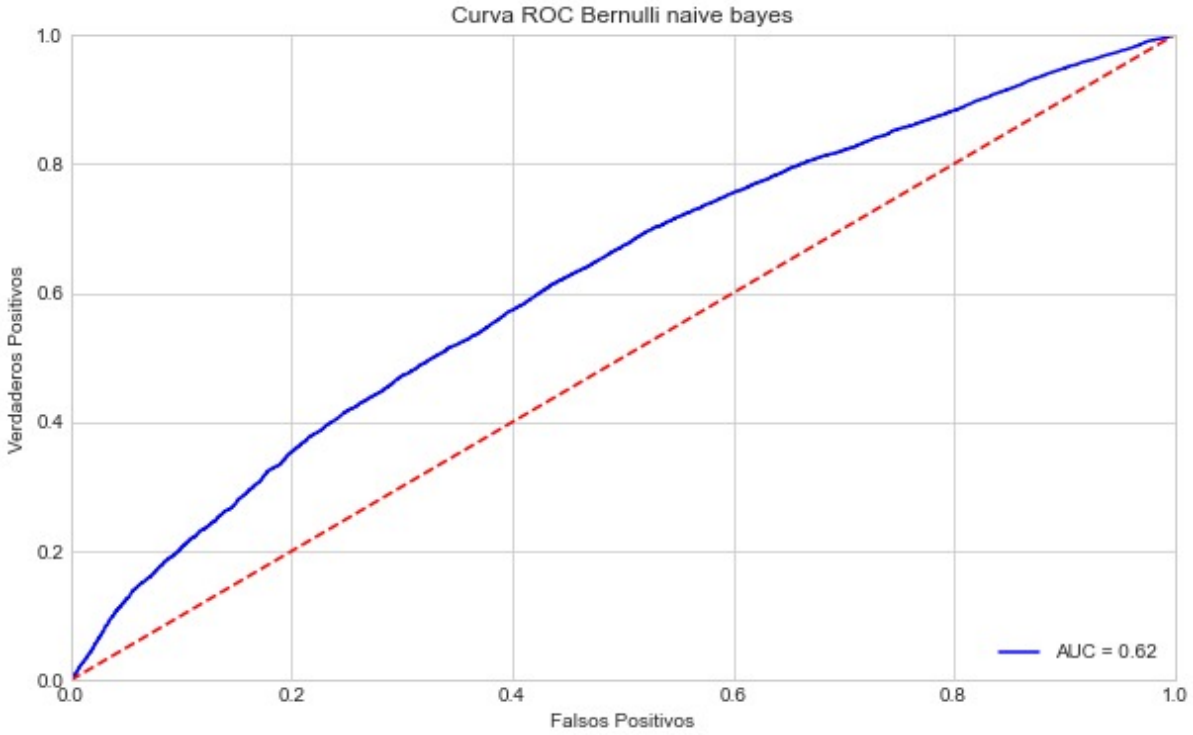


Para nuestro caso la curva auc es de 0,65 por lo que es posible aún mejorar la predicción de la variable objetivo.

2.3.2 - Naive Bayes

	precision	recall	f1-score	support
0	0.81	0.88	0.85	21976
1	0.34	0.23	0.28	5748
accuracy			0.75	27724
macro avg	0.58	0.56	0.56	27724
weighted avg	0.72	0.75	0.73	27724

El Modelo Bayes Bernoulli consigue una precisión de 72%
Para el f1, en los casos 0 es de 0.85 y para los casos en que el cliente no paga el crédito(1) es de 0.28.

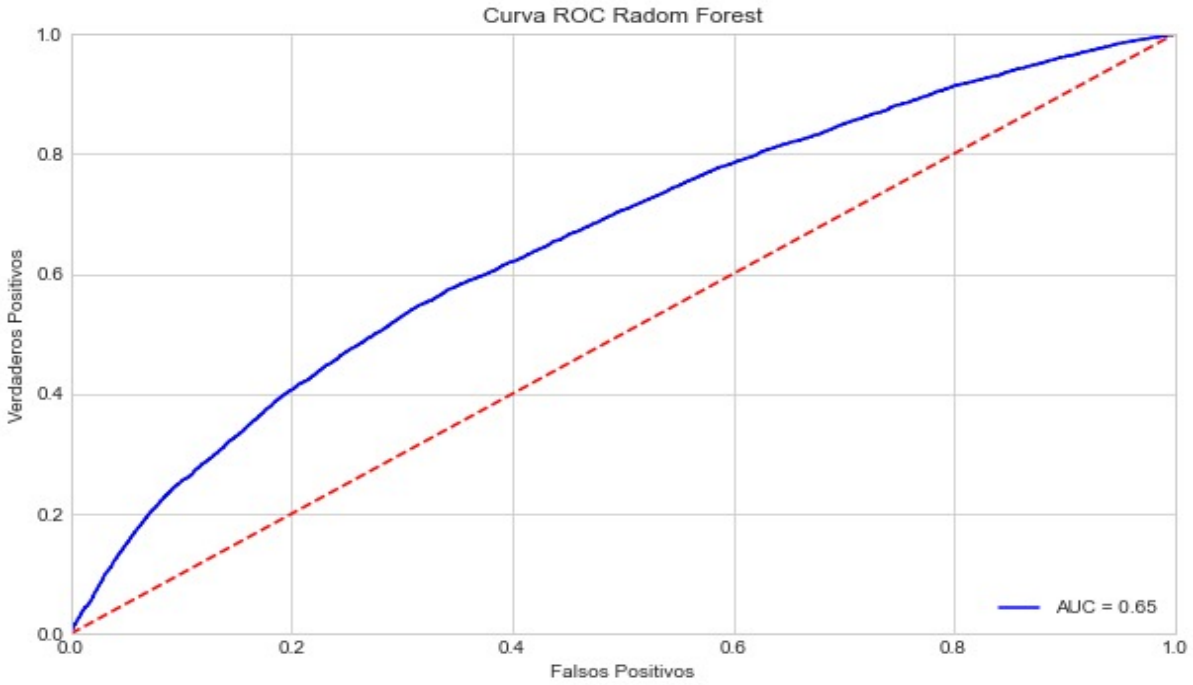


Para nuestro caso la curva auc es de 0,62 por lo que es posible aún mejorar la predicción de la variable objetivo.

2.3.3 - Random Forest

	precision	recall	f1-score	support
0	0.86	0.66	0.74	21976
1	0.30	0.57	0.40	5748
accuracy			0.64	27724
macro avg	0.58	0.62	0.57	27724
weighted avg	0.74	0.64	0.67	27724

La precisión del modelo Random Forest es del 74%.
Para el f1, en los casos 0 es de 0.74 y para los casos en que el cliente no paga el crédito(1) es de 0.40



Para nuestro caso la curva auc es de 0,65 por lo que es posible aún mejorar la predicción de la variable objetivo.

3.1 - Resultado de modelos candidatos.

Se probaron diferentes modelos candidatos para la problemática de explicar el buen y mal comportamiento futuro de los clientes de Consumo de una entidad bancaria.

De los diferentes experimentos realizados, se encontraron dos modelos que tienen las mejores métricas en comparación al resto de los modelos utilizados.

4.1 - Conclusiones.

Se estudiaron modelos con y sin hiperparametros encontrando mejores resultados con los modelos con hiperparametros.

- El modelo Logistic Regression
    - Consigue una precisión de general 74%.
    - Clase 0 obtiene un F1-Score de: 0,72
    - Clase 1 obtiene un F1-Score de: 0,39

- El modelo Random Forest
    - Consigue una precisión de general 74%.
    - Clase 0 obtiene un F1-Score de: 0,73
    - Clase 1 obtiene un F1-Score de: 0,40

Problemas:

- Los principales problemas se basaron en trabajar con una muestra desbalanceada.
- Problemas de convergencia al intentar generar un modelo descriptivo, desde la econometría utilizando regresión logística.

Propuestas de mejora:

- Buscar nuevas formas de preprocesamiento de la información.
- Mejorar el balanceo de la muestra.

Anexo 1 Valores perdidos

1.1 Tabla de datos perdidos

N	Variable	% Nulos	Acción
1	SK_ID_CURR	0%	Eliminar ID
2	TARGET	0%	
3	NAME_CONTRACT_TYPE	0%	Eliminar ID
4	CODE_GENDER	0%	
5	FLAG_OWN_CAR	0%	
6	FLAG_OWN_REALTY	0%	
7	CNT_CHILDREN	0%	
8	AMT_INCOME_TOTAL	0%	
9	AMT_CREDIT	0%	
10	AMT_ANNUITY	0%	
11	AMT_GOODS_PRICE	0%	
12	NAME_TYPE_SUITE	0%	Eliminar
13	NAME_INCOME_TYPE	0%	
14	NAME_EDUCATION_TYPE	0%	
15	NAME_FAMILY_STATUS	0%	
16	NAME_HOUSING_TYPE	0%	
17	REGION_POPULATION_RELATIVE	0%	
18	DAYS_BIRTH	0%	Se elimina la variable edad, debido a que edad tiene solo valores negativos y eso no es un dato válido
19	DAYS_EMPLOYED	0%	Se elimina la variable por tener valores no válidos(negativos) y valores que representa trabajar sobre 1000 años.
20	DAYS_REGISTRATION	0%	Se elimina por no tener sentido saber cuantos dias previos trabajo el cliente, si en este caso tenemos renta

21	DAYS_ID_PUBLISH	0%	Se elimina por no tener sentido saber cuantos dias previos trabajo el cliente a la última modificación de documentos
22	OWN_CAR_AGE	66%	Se elimina por no tener sentido saber cuantos dias previos trabajo el cliente a la última publicación de documentos
23	FLAG_MOBIL	0%	Variable eliminada por ser constante (solo un registro distinto)
24	FLAG_EMP_PHONE	0%	
25	FLAG_WORK_PHONE	0%	
26	FLAG_CONT_MOBILE	0%	Variable eliminada por ser constante (solo un registro distinto)
27	FLAG_PHONE	0%	
28	FLAG_EMAIL	0%	
29	OCCUPATION_TYPE	32%	Codifican los valores ausentes en una nueva categoría
30	CNT_FAM_MEMBERS	0%	
31	REGION_RATING_CLIENT	0%	
32	REGION_RATING_CLIENT_W_CITY	0%	
33	WEEKDAY_APPR_PROCESS_START	0%	se elimina porque no hace sentido incorporar el dia habil que el cliente solicita el crédito
34	HOUR_APPR_PROCESS_START	0%	se elimina porque no hace sentido incorporar la hora en que el cliente solicita el crédito
35	REG_REGION_NOT_LIVE_REGION	0%	
36	REG_REGION_NOT_WORK_REGION	0%	
37	LIVE_REGION_NOT_WORK_REGION	0%	
38	REG_CITY_NOT_LIVE_CITY	0%	
39	REG_CITY_NOT_WORK_CITY	0%	
40	LIVE_CITY_NOT_WORK_CITY	0%	
41	ORGANIZATION_TYPE	0%	
42	EXT_SOURCE_1	57%	Eliminar
43	EXT_SOURCE_2	0%	Se elimina por no tener información la descripción de la variable
44	EXT_SOURCE_3	20%	Se elimina por no tener información la descripción de la variable y es imposible con la información disponible imputar
45	APARTMENTS_AVG	51%	Se elimina por tener missing alto y por ser un promedio no se puede estimar de forma eficiente
46	BASEMENTAREA_AVG	59%	Se elimina por tener missing alto y por ser un promedio no se puede estimar de forma eficiente
47	YEARS_BEGINEXPLUATATION_AVG	49%	Se elimina por tener missing alto y por ser un promedio no se puede estimar de forma eficiente
48	YEARS_BUILD_AVG	67%	Se elimina por tener missing alto y por ser un promedio no se puede estimar de forma eficiente
49	COMMONAREA_AVG	70%	Se elimina por tener missing alto y por ser un promedio no se puede estimar de forma eficiente
50	ELEVATORS_AVG	54%	Se elimina por tener missing alto y por ser un promedio no se puede estimar de forma eficiente
51	ENTRANCES_AVG	51%	Se elimina por tener missing alto y por ser un promedio no se puede estimar de forma eficiente

52	FLOORSMAX_AVG	50%	Se elimina por tener missing alto y por ser un promedio no se puede estimar de forma eficiente
53	FLOORSMIN_AVG	68%	Se elimina por tener missing alto y por ser un promedio no se puede estimar de forma eficiente
54	LANDAREA_AVG	60%	Se elimina por tener missing alto y por ser un promedio no se puede estimar de forma eficiente
55	LIVINGAPARTMENTS_AVG	69%	Se elimina por tener missing alto y por ser un promedio no se puede estimar de forma eficiente
56	LIVINGAREA_AVG	51%	Se elimina por tener missing alto y por ser un promedio no se puede estimar de forma eficiente
57	NONLIVINGAPARTMENTS_AVG	70%	Eliminar
58	NONLIVINGAREA_AVG	56%	Eliminar
59	APARTMENTS_MODE	51%	Eliminar
60	BASEMENTAREA_MODE	59%	Eliminar
61	YEARS_BEGINEXPLUATATION_MODE	49%	Eliminar
62	YEARS_BUILD_MODE	67%	Eliminar
63	COMMONAREA_MODE	70%	Eliminar
64	ELEVATORS_MODE	54%	Eliminar
65	ENTRANCES_MODE	51%	Eliminar
66	FLOORSMAX_MODE	50%	Eliminar
67	FLOORSMIN_MODE	68%	Eliminar
68	LANDAREA_MODE	60%	Eliminar
69	LIVINGAPARTMENTS_MODE	69%	Eliminar
70	LIVINGAREA_MODE	51%	Eliminar
71	NONLIVINGAPARTMENTS_MODE	70%	Eliminar
72	NONLIVINGAREA_MODE	56%	Eliminar
73	APARTMENTS_MEDI	51%	Eliminar
74	BASEMENTAREA_MEDI	59%	Eliminar
75	YEARS_BEGINEXPLUATATION_MEDI	49%	Eliminar
76	YEARS_BUILD_MEDI	67%	Eliminar
77	COMMONAREA_MEDI	70%	Eliminar
78	ELEVATORS_MEDI	54%	Eliminar
79	ENTRANCES_MEDI	51%	Eliminar
80	FLOORSMAX_MEDI	50%	Eliminar
81	FLOORSMIN_MEDI	68%	Eliminar

82	LANDAREA_MEDI	60%	Eliminar
83	LIVINGAPARTMENTS_MEDI	69%	Eliminar
84	LIVINGAREA_MEDI	51%	Eliminar
85	NONLIVINGAPARTMENTS_MEDI	70%	Eliminar
86	NONLIVINGAREA_MEDI	56%	Eliminar
87	FONDKAPREMONT_MODE	69%	Eliminar
88	HOUSETYPE_MODE	51%	Eliminar
89	TOTALAREA_MODE	49%	Eliminar
90	WALLSMATERIAL_MODE	51%	Eliminar
91	EMERGENCYSTATE_MODE	48%	Eliminar
92	OBS_30_CNT_SOCIAL_CIRCLE	0%	
93	DEF_30_CNT_SOCIAL_CIRCLE	0%	
94	OBS_60_CNT_SOCIAL_CIRCLE	0%	
95	DEF_60_CNT_SOCIAL_CIRCLE	0%	
96	DAYS_LAST_PHONE_CHANGE	0%	
97	FLAG_DOCUMENT_2	0%	Se elimina por no tener informacion la descripcion de la variable
98	FLAG_DOCUMENT_3	0%	Se elimina por no tener informacion la descripcion de la variable
99	FLAG_DOCUMENT_4	0%	Se elimina por no tener informacion la descripcion de la variable
100	FLAG_DOCUMENT_5	0%	Se elimina por no tener informacion la descripcion de la variable
101	FLAG_DOCUMENT_6	0%	Se elimina por no tener informacion la descripcion de la variable
102	FLAG_DOCUMENT_7	0%	Se elimina por no tener informacion la descripcion de la variable
103	FLAG_DOCUMENT_8	0%	Se elimina por no tener informacion la descripcion de la variable
104	FLAG_DOCUMENT_9	0%	Se elimina por no tener informacion la descripcion de la variable
105	FLAG_DOCUMENT_10	0%	Se elimina por no tener informacion la descripcion de la variable
106	FLAG_DOCUMENT_11	0%	Se elimina por no tener informacion la descripcion de la variable
107	FLAG_DOCUMENT_12	0%	Se elimina por no tener informacion la descripcion de la variable
108	FLAG_DOCUMENT_13	0%	Se elimina por no tener informacion la descripcion de la variable
109	FLAG_DOCUMENT_14	0%	Se elimina por no tener informacion la descripcion de la variable
110	FLAG_DOCUMENT_15	0%	Se elimina por no tener informacion la descripcion de la variable
111	FLAG_DOCUMENT_16	0%	Se elimina por no tener informacion la descripcion de la variable
112	FLAG_DOCUMENT_17	0%	Se elimina por no tener informacion la descripcion de la variable
113	FLAG_DOCUMENT_18	0%	Se elimina por no tener informacion la descripcion de la variable

114	FLAG_DOCUMENT_19	0%	Se elimina por no tener informacion la descripcion de la variable
115	FLAG_DOCUMENT_20	0%	Se elimina por no tener informacion la descripcion de la variable
116	FLAG_DOCUMENT_21	0%	Se elimina por no tener informacion la descripcion de la variable
117	AMT_REQ_CREDIT_BUREAU_HOUR	13%	Se elimina porque no hace sentido tener las consultas al rut de una persona una hora antes de solicitar credito
118	AMT_REQ_CREDIT_BUREAU_DAY	13%	Se elimina porque no hace sentido tener las consultas al rut de una persona un dia antes de solicitar credito
119	AMT_REQ_CREDIT_BUREAU_WEEK	13%	Se elimina porque no hace sentido tener las consultas al rut de una persona una semana antes de solicitar credito
120	AMT_REQ_CREDIT_BUREAU_MON	13%	
121	AMT_REQ_CREDIT_BUREAU_QRT	13%	
122	AMT_REQ_CREDIT_BUREAU_YEAR	13%	

Anexo 2
Resultado modelos sin hiperparametros

2.1 - Regresión Logística

Regresion Logistica

In [39]:
```
lr =LogisticRegression().fit(X_train, y_train)
```

In [40]:
```
print(classification_report(y_test, lr.predict(X_test)))
```

```
              precision    recall  f1-score   support

           0       0.92      1.00      0.96     66531
           1       0.00      0.00      0.00      5764

    accuracy                           0.92     72295
   macro avg       0.46      0.50      0.48     72295
weighted avg       0.85      0.92      0.88     72295
```

El Modelo Regresión Logistica consigue una precisión de 85%

Para el f1, en los casos 0 es de 0.96 y para los casos en que el cliente se le no paga el crédito(1) es de 0.00

2.2 - Naive Bayes

Naive Bayes

In [41]:
```
bnb = BernoulliNB().fit(X_train, y_train)
```

In [42]:
```
print(classification_report(y_test, bnb.predict(X_test)))
```

```
              precision    recall  f1-score   support

           0       0.92      0.99      0.95     66531
           1       0.16      0.03      0.04      5764

    accuracy                           0.91     72295
   macro avg       0.54      0.51      0.50     72295
weighted avg       0.86      0.91      0.88     72295
```

El Modelo Bayes Bernoulli consigue una precisión de 86%
Para el f1, en los casos 0 es de 0.95 y para los casos en que el cliente no paga el crédito(1) es de 0.040.

2.3 - Random Forest

Random Forest

In [43]:
```
rfc = RandomForestClassifier(oob_score=True).fit(X_train, y_train)
```

In [45]:
```
print(classification_report(y_test, rfc.predict(X_test)))
```

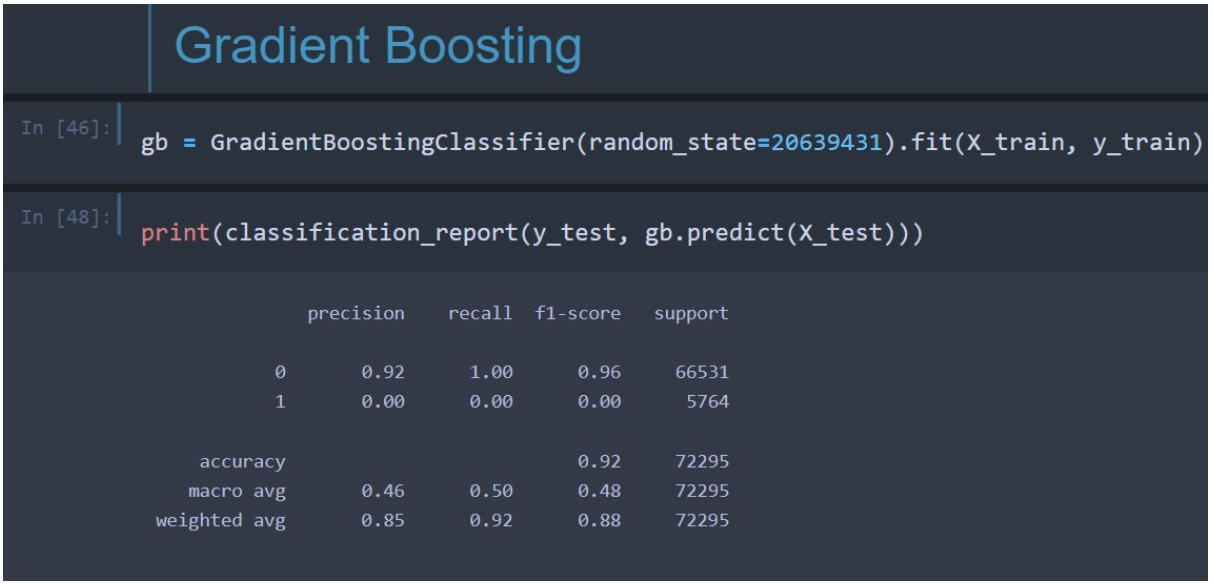
```
              precision    recall  f1-score   support

           0       0.92      1.00      0.96     66531
           1       1.00      0.00      0.00      5764

    accuracy                           0.92     72295
   macro avg       0.96      0.50      0.48     72295
weighted avg       0.93      0.92      0.88     72295
```

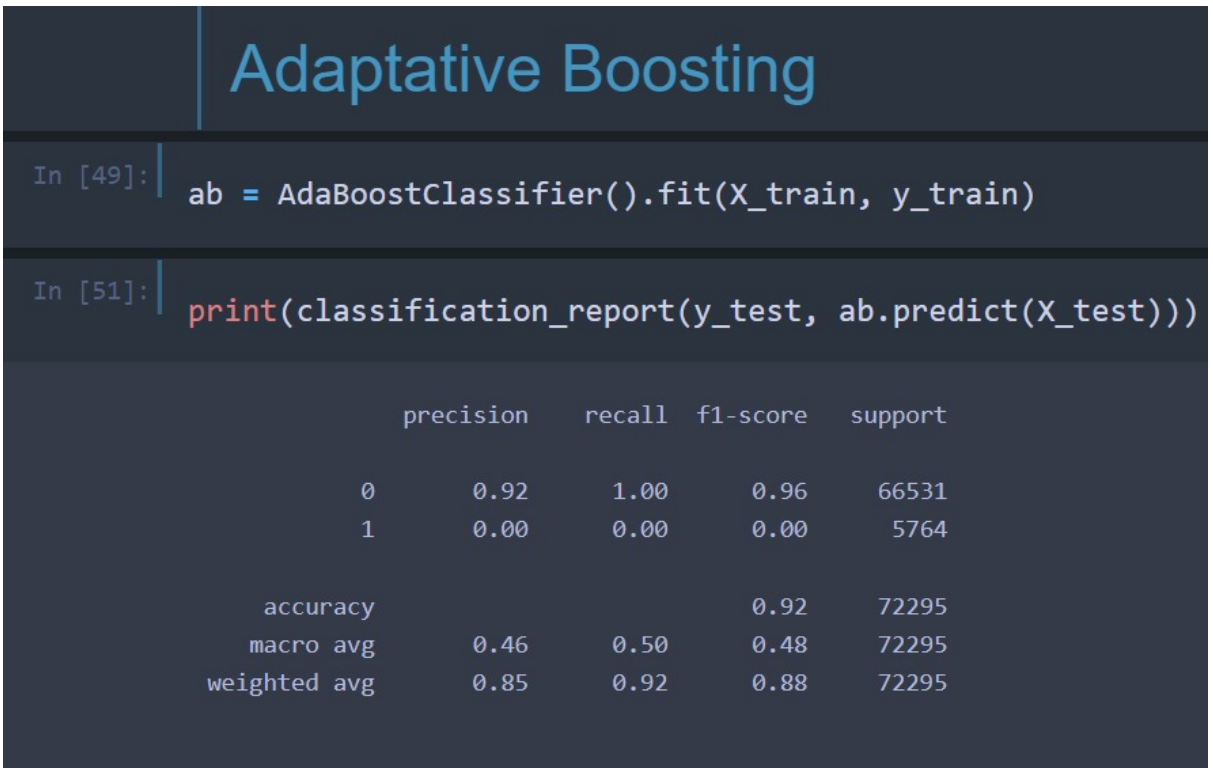
La precisión del modelo Random Forest es del 93%.
Para el f1, en los casos 0 es de 0.96 y para los casos en que el cliente no paga el crédito(1) es de 0.0

2.4 - Gradient Boosting



La precisión del modelo Gradient Boosting es del 85%. Para el f1, en los casos 0 es de 0.96 y para los casos en que el cliente no paga el crédito(1) es de 0.0

2.5 - Adaptative Boosting



La precisión del modelo Adaptative Boosting es del 85%. Para el f1, en los casos 0 es de 0.96 y para los casos en que el cliente no paga el crédito(1) es de 0.0

Anexo 3

Resultado modelos con hiperparametros y muestra seudo balanceada

3.1 - Gradient Boosting

	precision	recall	f1-score	support
0	0.80	0.99	0.88	21976
1	0.54	0.02	0.04	5748
accuracy			0.79	27724
macro avg	0.67	0.51	0.46	27724
weighted avg	0.74	0.79	0.71	27724

La precisión del modelo Gradient Boosting es del 74%. Para el f1, en los casos 0 es de 0.88 y para los casos en que el cliente no paga el crédito(1) es de 0.04

3.2 - Adaptative Boosting

	precision	recall	f1-score	support
0	0.79	1.00	0.88	21976
1	0.00	0.00	0.00	5748
accuracy			0.79	27724
macro avg	0.40	0.50	0.44	27724
weighted avg	0.63	0.79	0.70	27724

La precisión del modelo Adaptative Boosting es del 63%. Para el f1, en los casos 0 es de 0.88 y para los casos en que el cliente no paga el crédito(1) es de 0.0