DETERMINACIÓN DE ALGORITMOS DE CLASIFICACIÓN ÓPTIMOS PARA LA EVALUACIÓN DE RIESGO CREDITICIO EN EL CASO DE PYMES QUE OPERAN EN PLATAFORMAS P2P

CURSO DATA SCIENCE
CODERHOUSE
MARZO 2023

JUAN GABRIEL GARCIA OJEDA
JUAN IGNACIO CARRERE

INDICE

CONTENIDO

PROBLEMÁTICA

Las instituciones financieras han visto incrementada la demanda de créditos a lo largo de las últimas décadas, siendo este aumento explicado en parte por el surgimiento de nuevos modelos de negocios dentro del sistema financiero comúnmente denominados finanzas descentralizadas. Esta situación, en conjunto con las diversas regulaciones que exigen los países con relación a la necesidad de llevar a cabo algún tipo de evaluación previa que determine el riesgo que implica el otorgamiento del crédito a determinado individuo o empresa, han llevado a dichas instituciones a desarrollar diversas técnicas que sirvan para estimar la probabilidad de default de la manera más eficiente y precisa posible. en este sentido, los modelos de puntaje crediticio han sido los más utilizados.

El surgimiento de fintechs como Peer to peer lending (en adelante P2P) ha puesto en mayor relieve la necesidad de mejorar las técnicas utilizadas en la estimación de los modelos de puntaje debido a las características propias de estas plataformas, ya que suponen, entre otras cosas, un riesgo sistémico mayor. Además, a medida que crece el volumen de operaciones a través de estas plataformas, una estimación inadecuada del riesgo crediticio puede implicar una amenaza para la estabilidad financiera.

En los últimos años se ha visto un avance importante en el uso de técnicas de Machine Learning para el desarrollo de modelos de puntaje crediticio como alternativa al uso de modelos econométricos clásicos como la Regresión Lineal y la Regresión Logística, entre otros. En ciertos contextos, los algoritmos de clasificación desarrollados a partir de Inteligencia Artificial han probado tener un mejor desempeño.

PREGUNTAS Y OBJETIVOS

-) **Pregunta general:** ¿Cuál es el algoritmo de clasificación de Machine Learning más adecuado para la estimación de modelos de puntaje crediticio aplicado a PyMES que participan en plataformas P2P?

-) **Pregunta específica 1:** ¿Puede cambiar la decisión del algoritmo óptimo obtenido a partir de la métrica "exactitud" al incorporar nuevas métricas?

-) **Pregunta específica 2:** ¿Es mejor el desempeño predictivo del algoritmo óptimo al del modelo Logit?

En este trabajo se busca llevar a cabo un análisis comparativo de los principales algoritmos de clasificación utilizados en la actualidad aplicados a la estimación de modelos de puntaje crediticio para la evaluación del riesgo de empresas PYME que acceden a financiamiento mediante instituciones no tradicionales, como la FinTech P2P. Se propone llevar a cabo dicho análisis a través del uso de diversas métricas que permitan obtener conclusiones e identificar ventajas y desventajas en cada uno de los algoritmos.

DATA ACQUISITION

El data set con el que trabajaremos esta provisto por la European External Credit Assessment Institution, el cual está compuesto por 23 columnas y 4515 filas. La base de datos completa se puede descargar en el siguiente enlace https://www.frontiersin.org/articles/10.3389/frai.2019.00003/full#supplementary-material.

Detalle de cada columna:

- **RATIO001** = (Total assets - Shareholders Funds)/Shareholders Funds): Capital ajeno en relacion al propio ó cociente de participación de tercero.
- **RATIO002** = (Long term debt + Loans)/Shareholders Funds): Obligaciones de Largo Plazo (pasivo no corriente) en relación al capital propio.

- **RATIO003** = (Total assets/Total liabilties): Ratio de solvencia. En general es solvente si es > 1.
- **RATIO004** = (Current assets/Current liabilties): Ratio de liquidez. Activos corrientes en relación con los pasivos corrientes.
- **RATIO005** = ((Current assets - Current assets: stocks)/Current liabilties)): Comúnmente llamado "Prueba ácida". Ratio de liquidez con los activos más líquidos.
- **RATIO006** = ((Shareholders Funds + Non current liabilities)/Fixed assets)): Indica la proporción en la que los activos de Largo Plazo son financiados por fuentes de igual o mayor plazo.
- **RATIO008** = (EBIT/interest paid) (EBIT = ganancias antes de intereses e impuestos): Capacidad para atender los intereses que paga la empresa.
- **RATIO011** = (Profit (loss) before tax + Interest paid)/Total assets)): Rendimiento del activo de la empresa sin considerar el pago de impuestos.
- **RATIO012** = ((P/L) after tax/Shareholders Funds): Ratio de rentabilidad de los accionistas después de impuestos.
- **RATIO017** = (Operating revenues/Total assets): Pesos de ganancia por actividad operativa por peso de financiamiento total de la empresa.
- **RATIO018** = (Sales/Total assets): Pesos de ingreso por ventas por peso de financiamiento total.
- **RATIO019** = (Interest paid/(Profit before taxes + Interest paid)): Peso de la carga financiera sobre el resultado antes de impuestos.
- **RATIO027** = (EBITDA/interest paid): Ganancia en términos de los intereses pagados
- **RATIO029** = (EBITDA/Operating revenues): Ganancia en términos de la ganancia operativa
- **RATIO030** = (EBITDA/Sales): Ganancia por venta.

- **DPO** = (Trade Payables/Operating revenues): Acreedores/Resultado operativo: Una forma de ver la solvencia, por cada peso de ganancia cuantos pesos debo.
- **DSO** = (Trade Receivables/Operating revenues): Cuentas por cobrar/Resultado operativo: Misma interpretación anterior, pero en este caso es cuantos pesos me deben
- **DIO** = (Inventories/Operating revenues): Valor del inventario en relación al resultado operativo
- **NACE** = (Industry classification on NACE code, 4 digits precision): identificación de industria en la que opera la empresa
- **STATUS** = 1 para default, 0 no default

También está incluida la columna 'Unnamed: 0' es ordinal y se define como índice para la base de datos.

DATA WRANGLING

Consiste en la manipulación, limpieza y unificación de conjuntos de datos complejos y desordenados para facilitar su acceso, análisis y modelado. El proceso incluye convertir y mapear los datos crudos, y dejarlos en un formato más adecuado para su uso.

INFORMACION Y CARACTERISTICAS DE LOS DATOS

Observamos las características de las columnas y cantidad de nulos

#	Column	Non-Null Count	Dtype
0	Unnamed: 0	4514 non-null	int64
1	X1	4514 non-null	int64
2	ratio001	4514 non-null	float64
3	ratio002	4514 non-null	float64
4	ratio003	4514 non-null	float64
5	ratio004	4514 non-null	float64
6	ratio005	4514 non-null	float64
7	ratio006	4514 non-null	float64
8	ratio008	4514 non-null	float64
9	ratio011	4514 non-null	float64
10	ratio012	4514 non-null	float64
11	ratio017	4514 non-null	float64
12	ratio018	4514 non-null	float64
13	ratio019	4514 non-null	float64
14	ratio027	4514 non-null	float64
15	ratio029	4514 non-null	float64
16	ratio030	4514 non-null	float64
17	DIO	4514 non-null	int64
18	DPO	4514 non-null	int64
19	DSO	4514 non-null	int64
20	turnover	4514 non-null	int64
21	status	4514 non-null	int64
22	nace	4514 non-null	int64

Las variables 'turnover' y 'X1' no presenta explicación por parte de la fuente de la base de datos y la variable 'nace' es un numero identificador de industria, por lo que se decide eliminar estas variables de la base.
Analizamos también la presencia o no de datos faltantes en la base de datos.

```
ratio001    0
ratio002    0
ratio003    0
ratio004    0
ratio005    0
ratio006    0
ratio008    0
ratio011    0
ratio012    0
ratio017    0
ratio018    0
ratio019    0
ratio027    0
ratio029    0
ratio030    0
DIO         0
DPO         0
DSO         0
status      0
dtype: int64
```

No se observa la presencia de datos nulos en la base de datos.
Buscamos identificar variables numéricas y variables categóricas. Para esto, se lleva a cabo un recuento de cantidad de valores únicos de cada columna de la base de datos. Para las variables dummy este recuento debería ser de 2 (0 y 1).
Podemos identificar rápidamente que la columna Status corresponde a valores binarios

```
Valores únicos por columna:
ratio001    2098
ratio002     874
ratio003     400
ratio004     513
ratio005     449
ratio006    1521
ratio008    2775
ratio011     142
ratio012     383
ratio017     487
ratio018     495
ratio019     334
ratio027    3068
ratio029     186
ratio030     198
DIO          522
DPO          406
DSO          471
status         2
dtype: int64
```

mientras que el resto de las variables son continuas.

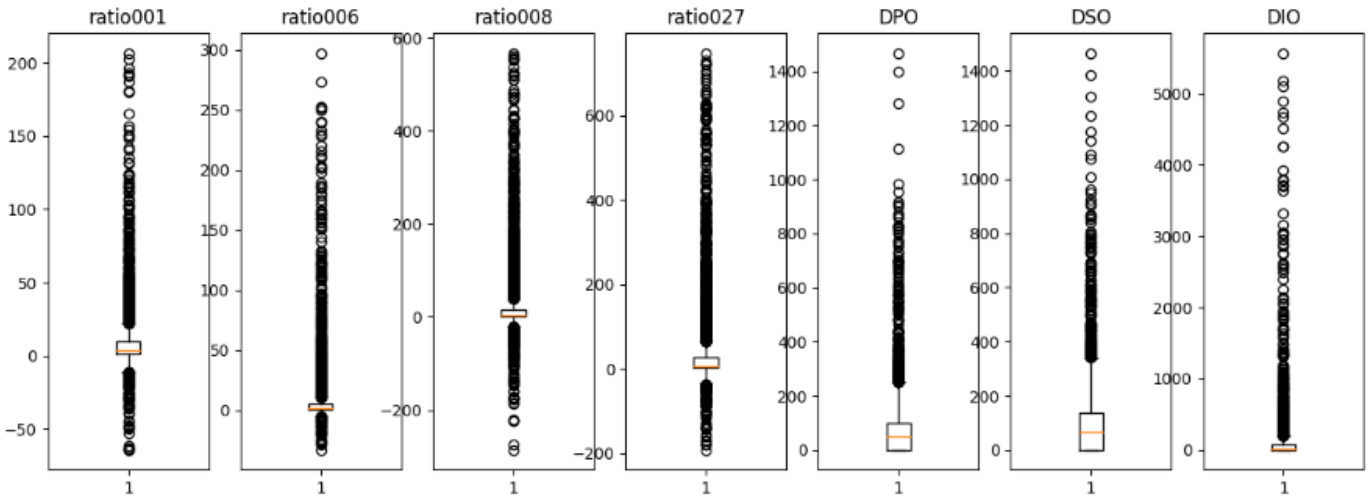
Resulta interesante analizar brevemente simples estadísticas de aquellas variables continuas, y mediante la función describe podemos tener una visión global de los datos numéricos.

	count	mean	std	min	25%	50%	75%	max
ratio001	4514.0	8.89	19.15	-64.43	1.30	3.77	9.68	206.55
ratio002	4514.0	1.26	3.33	-9.58	0.00	0.09	1.17	33.38
ratio003	4514.0	1.44	0.76	0.17	1.07	1.20	1.52	8.27
ratio004	4514.0	1.54	1.20	0.01	0.97	1.22	1.72	13.71
ratio005	4514.0	1.19	1.02	0.00	0.61	0.99	1.41	10.88
ratio006	4514.0	7.93	23.15	0.00	0.94	1.72	4.89	297.02
ratio008	4514.0	23.07	70.27	-285.86	1.24	3.59	16.32	566.96
ratio011	4514.0	0.03	0.15	-1.28	0.01	0.03	0.07	0.49
ratio012	4514.0	-0.07	0.79	-8.54	0.00	0.07	0.21	1.08
ratio017	4514.0	1.37	1.07	0.01	0.68	1.17	1.74	8.42
ratio018	4514.0	1.34	1.06	0.01	0.64	1.13	1.70	8.42
ratio019	4514.0	0.19	0.50	-3.32	0.01	0.10	0.39	3.95
ratio027	4514.0	36.51	92.89	-191.63	2.47	7.30	27.61	747.01
ratio029	4514.0	0.06	0.20	-2.08	0.02	0.06	0.11	0.94
ratio030	4514.0	0.07	0.22	-2.39	0.02	0.06	0.12	1.28
DIO	4514.0	105.23	355.81	0.00	1.00	19.00	80.00	5569.00
DPO	4514.0	74.81	105.29	0.00	0.00	51.00	99.00	983.00
DSO	4514.0	93.45	118.16	0.00	0.00	67.00	136.00	960.00

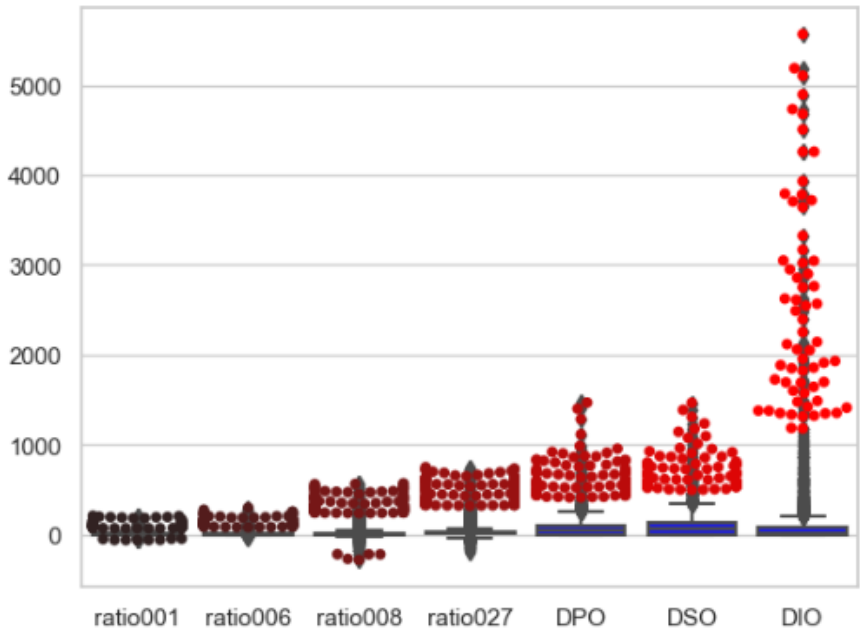
Teniendo en cuenta que las variables analizadas son ratios financieros, y por lo tanto es esperable que tengan cierta amplitud podemos notar que observando los valores minimos y máximos de cada variable vemos que los siguientes ratios presentan bastante amplitud:

* RATIO001
* RATIO006
* RATIO008
* RATIO027
* DIO
* DPO
* DSO

Analizamos estas variables mediante gráficos de boxplot



Los puntos negros fuera de la 'caja' son considerados outliers. Podemos visualizarlos mejor mediante el siguiente gráfico en el cual se destacan dichos valores.



De los gráficos de boxplot de las variables RatioXXX podemos concluir:

* Del **RATIO001** (ACTIVO TOTAL - CAPITAL PROPIO/CAPITAL PROPIO): Es posible que una empresa PyME se vea financiado por capital ajeno por hasta un 200% de su capital propio.

Esto significa que la empresa está altamente apalancada. Por otro lado, que el ratio sea negativo implica que el financiamiento propio es mayor proporcionalmente al financiamiento ajeno. Este ratio depende de la estructura de financiamiento que haya decidido utilizar la empresa.

* Del **RATIO006** (CAPITAL PROPIO + PASIVOS NO CORRIENTES/ACTIVOS FIJOS): Es posible que una empresa PyME tenga financiamiento de hasta 3 veces de los activos totales con fuentes de igual o mayor plazo (es más, es deseable en ciertos contextos que sea así). Por otro lado, valores negativos del ratio no son posibles, o por lo menos no es esperable y probablemente sea un error en la carga de datos

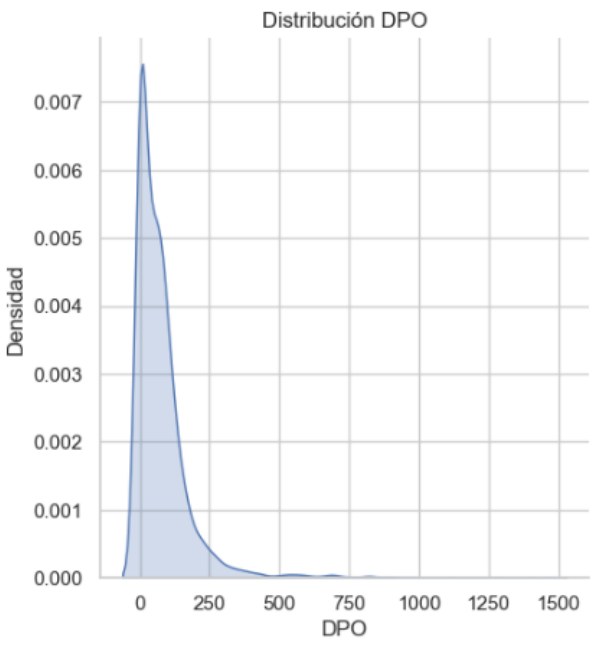
* Del **RATIO008** (EBIT/INTERESES PAGADOS): Es posible, y probable, que este ratio alcance valores altos ya que en una empresa con poca carga de intereses y buen nivel de resultado implicaría que su EBIT pueda llegar a ser 6 veces la carga de intereses. Los valores negativos implicarían EBIT negativo (resultado antes de impuestos e intereses negativo = pérdida)

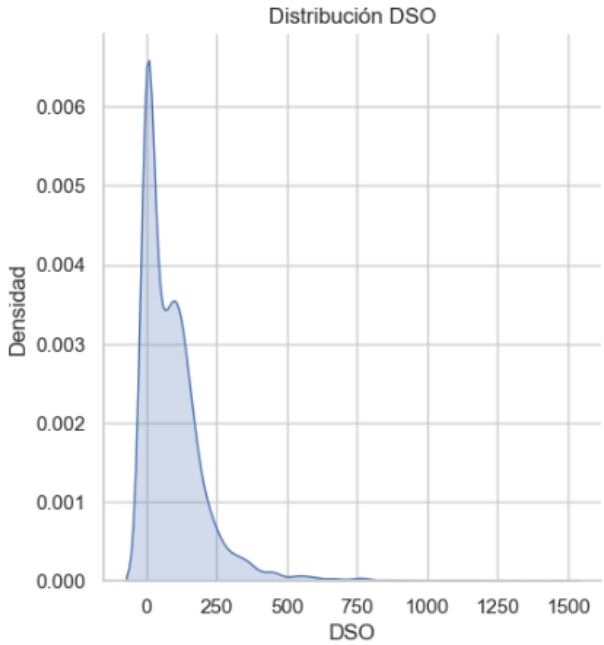
* Del **RATIO027** (EBITDA/INTERESES PAGADOS): Misma explicación que para el RATIO008, solo que al resultado se le restan además las depreciaciones y amortizaciones.

De las conclusiones se decide eliminar los valores negativos del RATIO006 y reemplazarlos por el valor 0 (que es el que implicaría un valor neutro en el ratio). Se decide dejar el resto de "outliers" debido a que representaría una pérdida de valiosa información que caracteriza a cada empresa en particular. Además, al ser ratios financieros estos datos son correspondidos con otras variables, por lo que la modificación arbitraria podría llevar a inconsistencias en la totalidad de la base de datos.

Para el caso de las variables DIO, DPO y DSO:

* Las variables **DPO** (ACREEDORES/GANANCIA OPERATIVA) y DSO (CUENTAS POR COBRAR/GANANCIA OPERATIVA) representan la situación de la empresa en la gestión de pagos y cobros. Al ser acreedores y cuentas por cobrar únicamente utilizados en estos dos ratios, de tratarse los valores atípicos no se perdería la correlación con otro ratio del resto de la base de datos. Sin embargo, únicamente se tratarán aquellos outliers que superen el valor de 1000, es decir, empresas que deban y/o les deban más de 10 veces su ganancia operativa lo cual es considerado inusual y puede deberse a un error en la carga de datos. Se analizará la distribución de ambas variables para determinar por cual valor de tendencia central es conveniente reemplazarlo.

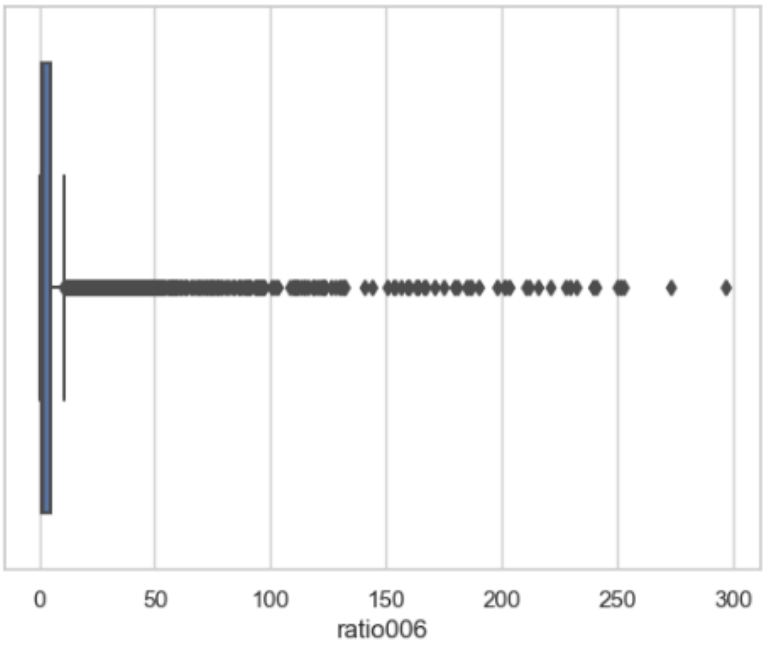


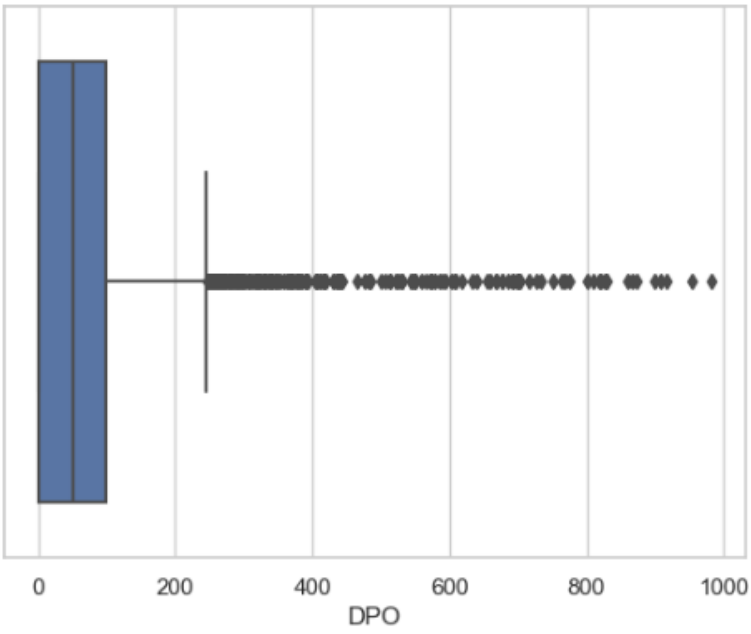


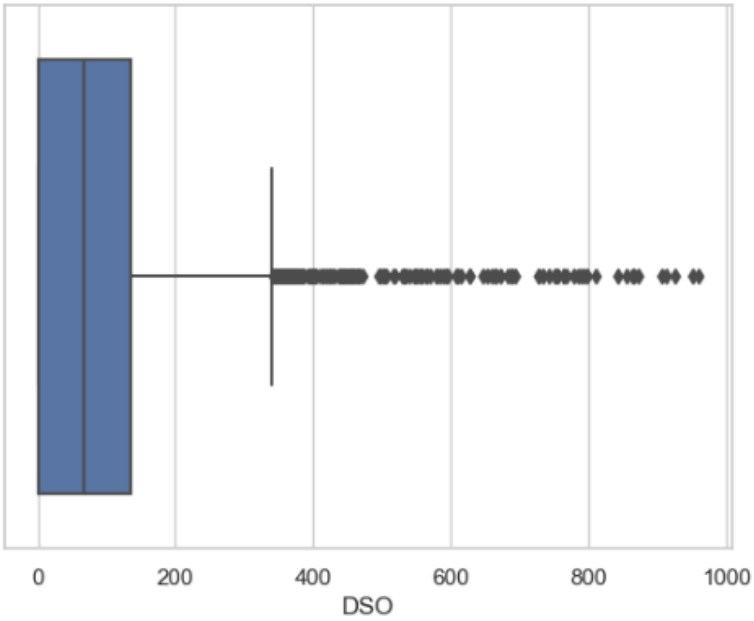
Se puede notar lo asimétrica que es la distribución de ambas variables, por lo que será óptimo reemplazar los outliers considerados por la mediana de la distribución.

* La variable **DIO** (Inventarios/Ganancia Operativa) se dejará con sus outliers, ya que el nivel de inventario de las empresas puede depender de la actividad que realice y cambiar dicho valor puede provocar perdida de valiosa información.

GRAFICOS BOXPLOT FINALES DE AQUELLAS VARIABLES MODIFICADAS







CONCLUSIONES DATA WRANGLING

* El DataSet con el que trabajaremos está compuesto por 4514 filas y 19 columnas (luego de eliminar 3 columnas sin ninguna utilidad y definir una de ellas como índice). Cada columna indica algún tipo de información financiera.

* No se detectan datos faltantes en la base de datos.

* 16 variables son numéricas continuas y solo 1 variable dicotómica (binaria).

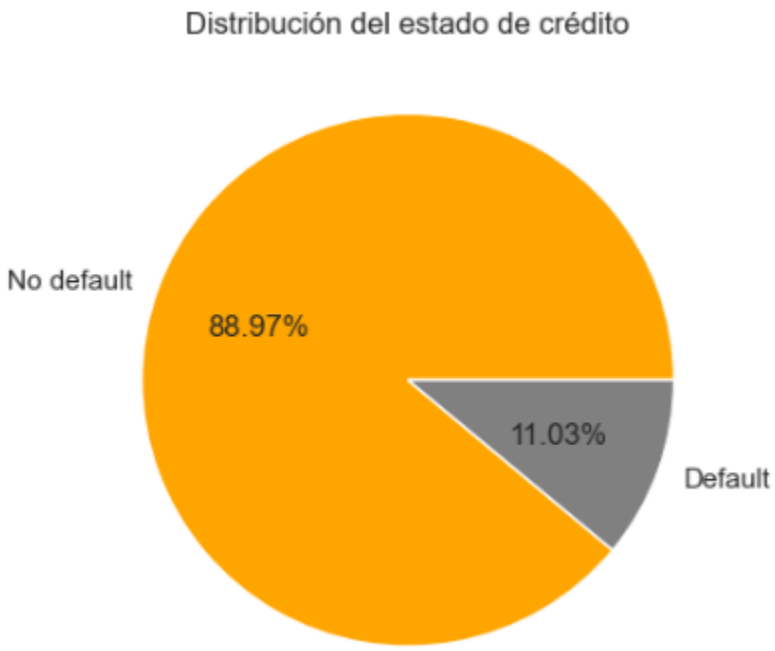
* De las variables que mayor amplitud presentan se buscó mediante grafico boxplot la presencia de datos atípicos. Se decidió dejar los valores que forman parte de los extremos de la distribución de cada variable, a excepción del caso de las variables 'DPO' y 'DSO'. La razón para decidir quedarse con los valores considerados atípicos según el método zscore es que no son considerados "errores" ya que es información valiosa sobre las características de cada empresa.

* Del ratio006 se eliminaron los valores negativos, ya que observando la conformación del ratio no es posible que surjan valores menores a cero.

* Se reemplazaron los valores mayores a 1000 para los casos de las variables 'DPO' y 'DSO'.

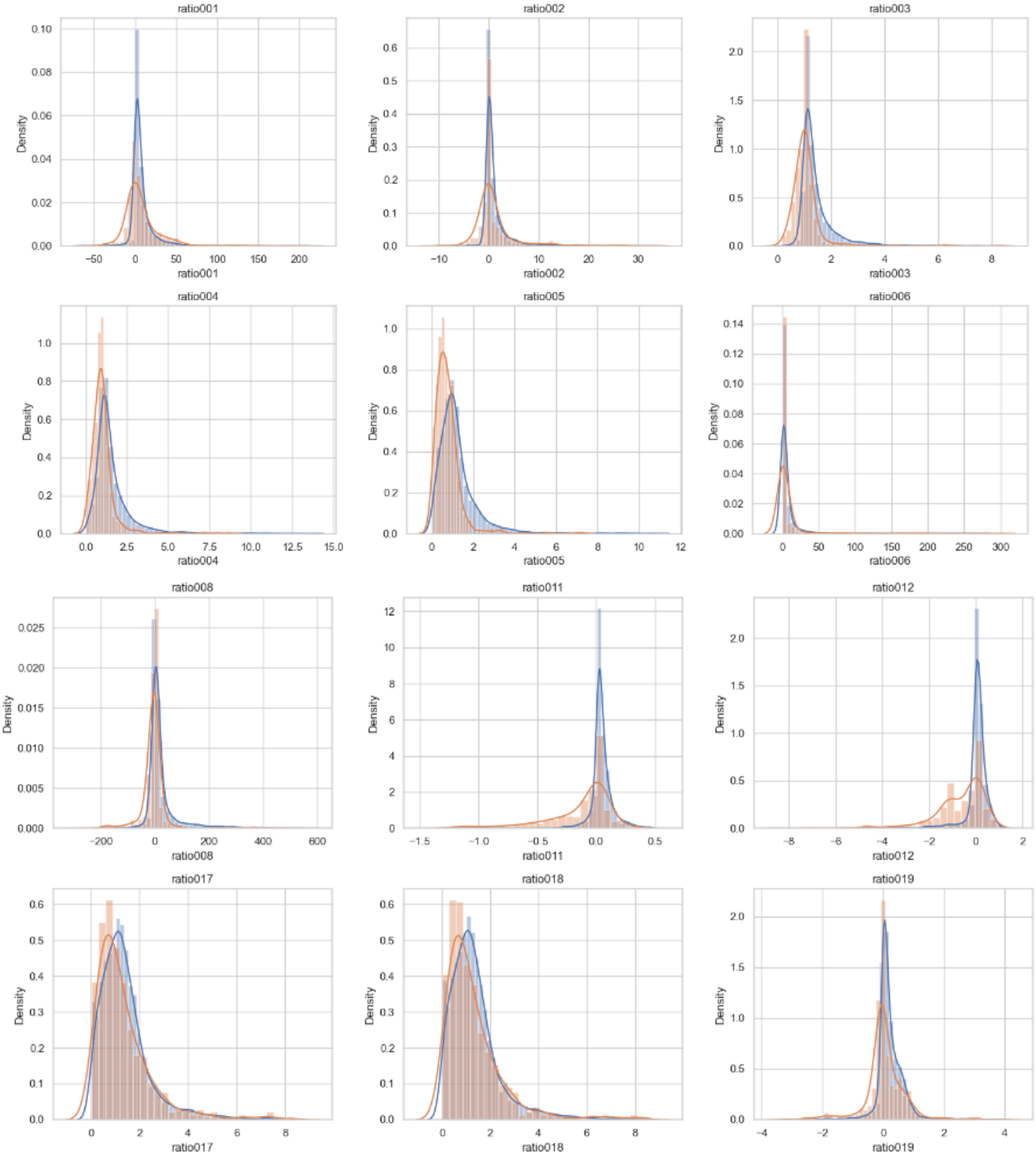
EXPLORATORY DATA ANALYSIS

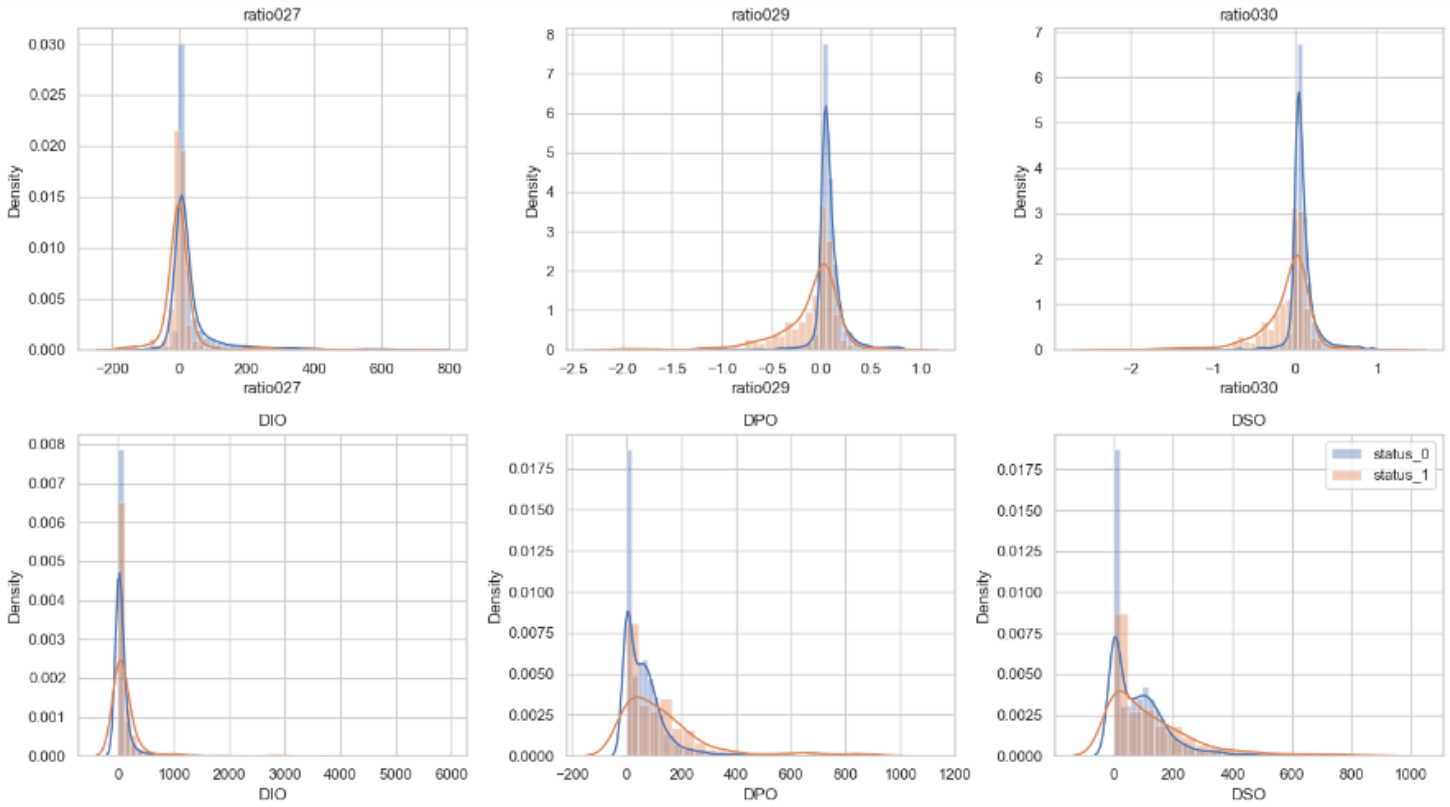
Comenzamos analizando la proporción de instancias en default (Status = 1) y el resto en ‘no default’ (Status = 0).



Del grafico de tortas podemos notar que la base de datos es desbalanceada. Mientras que los datos clasificados como 'No default' o 0 representan el 88.97% del total, aquellos datos clasificados como 'Default' o 1 representan el 11.03% del total.

Es interesante observar las diferentes características de aquellas empresas que cumplen con el pago de sus obligaciones respecto a las que no. Para ello, graficamos un ‘distplot’ de cada una de las variables numéricas del dataset.


ratio001
Density
ratio001
ratio002
Density
ratio002
ratio003
Density
ratio003
ratio004
Density
ratio004
ratio005
Density
ratio005
ratio006
Density
ratio006
ratio008
Density
ratio008
ratio011
Density
ratio011
ratio012
Density
ratio012
ratio017
Density
ratio017
ratio018
Density
ratio018
ratio019
Density
ratio019



Mediante un test de medias podemos definir si existe una diferencia estadística entre las medias de los datos correspondientes a cada clase para cada ratio. El p-valor que resultante de cada test indica si la diferencia entre medias de cada clase es estadísticamente significativa o no. A partir de un p-valor menor a 0.05 se considera a la diferencia estadísticamente significativa.

El test de medias de los siguientes ratios resultaron estadísticamente significativos

status	ratio003	ratio004	ratio005	ratio008	ratio011	ratio012	ratio019
0	1.487742	1.597079	1.243633	26.217702	0.047816	0.008752	0.211768
1	1.086767	1.041546	0.757088	-2.334116	-0.133996	-0.694699	0.050361

status	ratio027	ratio029	ratio030	DIO	DPO	DSO
0	40.178063	0.084345	0.091696	100.609313	66.687998	89.090388
1	6.955663	-0.118976	-0.119036	142.471888	140.339357	128.646586

A partir de los gráficos de las distribuciones y del resultado del test de medias podemos llegar a las siguientes conclusiones (recordando que 1 indica default):

- **ratio003**: en promedio el ratio de solvencia (AT/PT) es mayor para la clase 0 (1.48) que para la clase 1 (1.08). Un menor valor de este ratio implica una menor solvencia.

- **ratio004**: en promedio el ratio de liquidez (AC/PC) es mayor para la clase 0 (1.59) que para la clase 1 (1.04). Un menor valor de este ratio implica una menor liquidez.

- **ratio005**: en promedio el ratio de liquidez extremo (AC - STOCKS/PC) es mayor para la clase 0 (1.24) que para la clase 1 (0.75). Un menor valor de este ratio implica una menor liquidez extrema.

- **ratio008**: en promedio el ratio (EBIT/Int. pagados) es mayor para la clase 0 (26.2) que para la clase 1 (-2.3). Un menor valor de este ratio implica una menor ganancia antes de intereses e impuestos en relación a los intereses, y un valor menor que cero implica EBIT negativo, es decir, pérdida.

- **ratio011**: en promedio el ratio de peso de la carga impositiva (P/L + Int. pagados/AT) es mayor para la clase 0 (0.04) que para la clase 1 (-0.13). Un menor valor de este ratio implica menor rendimiento del activo total antes de impuestos, y un valor menor que cero implica rendimiento negativo del activo total antes de impuestos.

- **ratio012**: en promedio el ratio de rentabilidad de los accionistas después de impuestos (P/L después de imp/Capital propio) es mayor para la clase 0 (0.008) que para la clase 1 (-0.69). Un menor valor de este ratio implica una menor rentabilidad para los accionistas luego de impuestos, y un valor menor que cero implica rentabilidad negativa después de impuestos para los accionistas.

- **ratio019**: en promedio el ratio de peso de la carga financiera sobre el resultado antes de impuestos (Int pagados/ P/L + Int pagados) es mayor para la clase 0 (0.21) que para la clase 1 (0.05). Un menor valor de este ratio implica una menor carga financiera (intereses) sobre el resultado antes de impuestos. Es esperable que el valor del ratio sea menor en empresas en default ya que la carga financiera baja al no cumplir con la totalidad de sus obligaciones.

- **ratio027**: en promedio el ratio (EBITDA/Int. pagados) es mayor para la clase 0 (40.17) que para la clase 1 (6.9). Un menor valor de este ratio implica una menor ganancia antes de intereses, impuestos, amortización y depreciaciones en relación a los intereses pagados.

- **ratio029**: en promedio el ratio (EBITDA/Resultado operativo) es mayor para la clase 0 (0.08) que para la clase 1 (-0.11). Un menor valor de este ratio implica una menor ganancia antes de intereses, impuestos, amortización y depreciaciones en relación al resultado operativo, y un valor menor que cero implica un EBITDA negativo promedio, es decir, pérdida promedio.

- **ratio030**: en promedio el ratio (EBITDA/Ventas) es mayor para la clase 0 (0.09) que para la clase 1 (-0.11). Un menor valor de este ratio implica una menor ganancia antes de intereses, impuestos, amortización y depreciaciones en relación a las ventas, y un valor menor que cero implica un EBITDA negativo promedio, es decir, pérdida promedio.

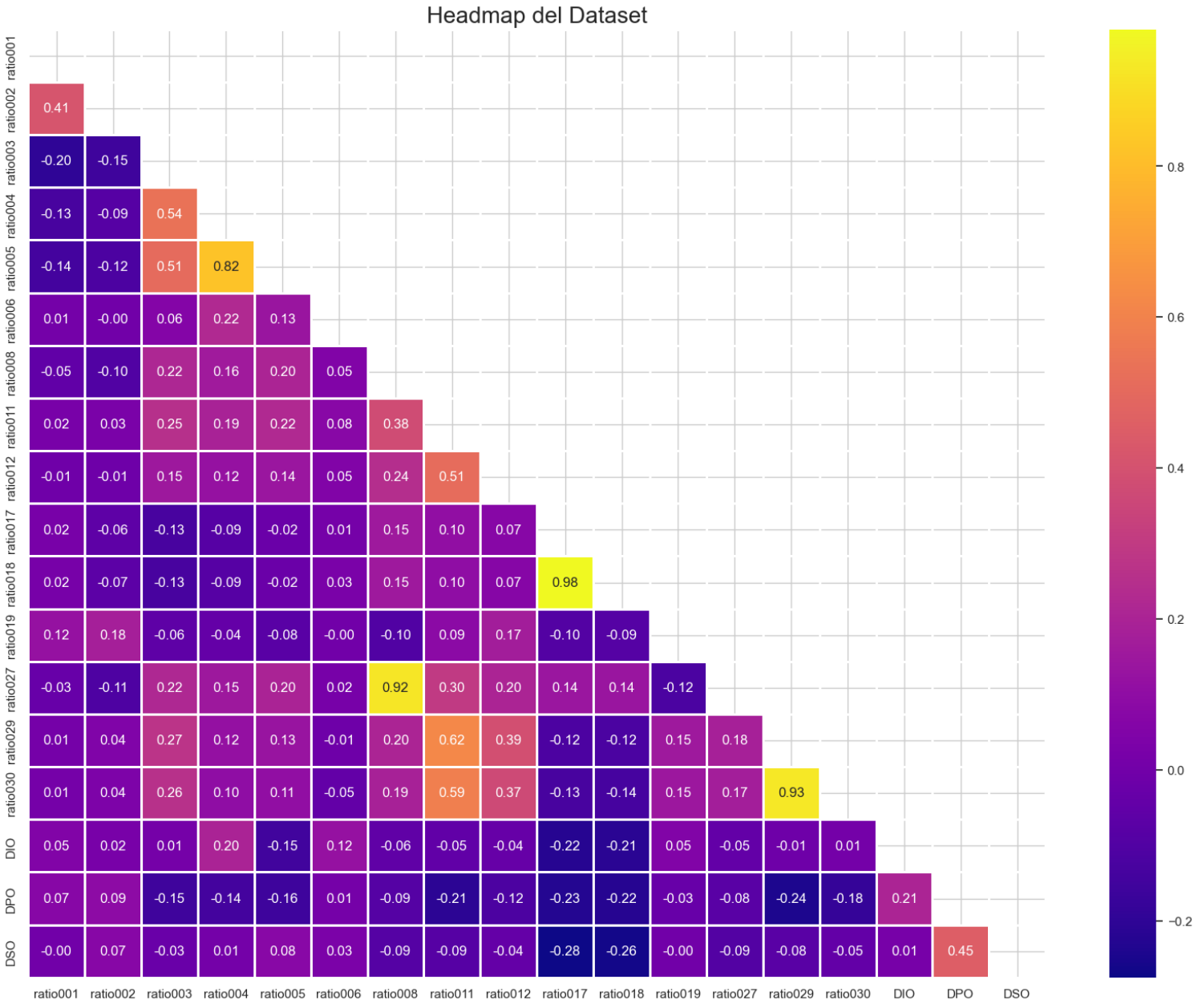
- **DIO**: en promedio el ratio (Inventario/Resultado operativo) es menor para la clase 0 (100.6) que para la clase 1 (142.4). Un menor valor de este ratio implica un menor valor de inventario en relación con el resultado operativo, lo que implica mayor rotación de mercadería.

- **DPO**: en promedio el ratio (Acreedores/Resultado operativo) es menor para la clase 0 (66.4) que para la clase 1 (140.2). Un menor valor de este ratio implica menor acreencias por parte de la empresa en relación al resultado operativo.

- **DSO**: en promedio el ratio (Cuentas por cobrar/Resultado operativo) es menor para la clase 0 (89.09) que para la clase 1 (128.6). Un menor valor de este ratio implica menor deudores por parte de la empresa en relación al resultado operativo.

Análisis de correlación entre variables numéricas del DataSet

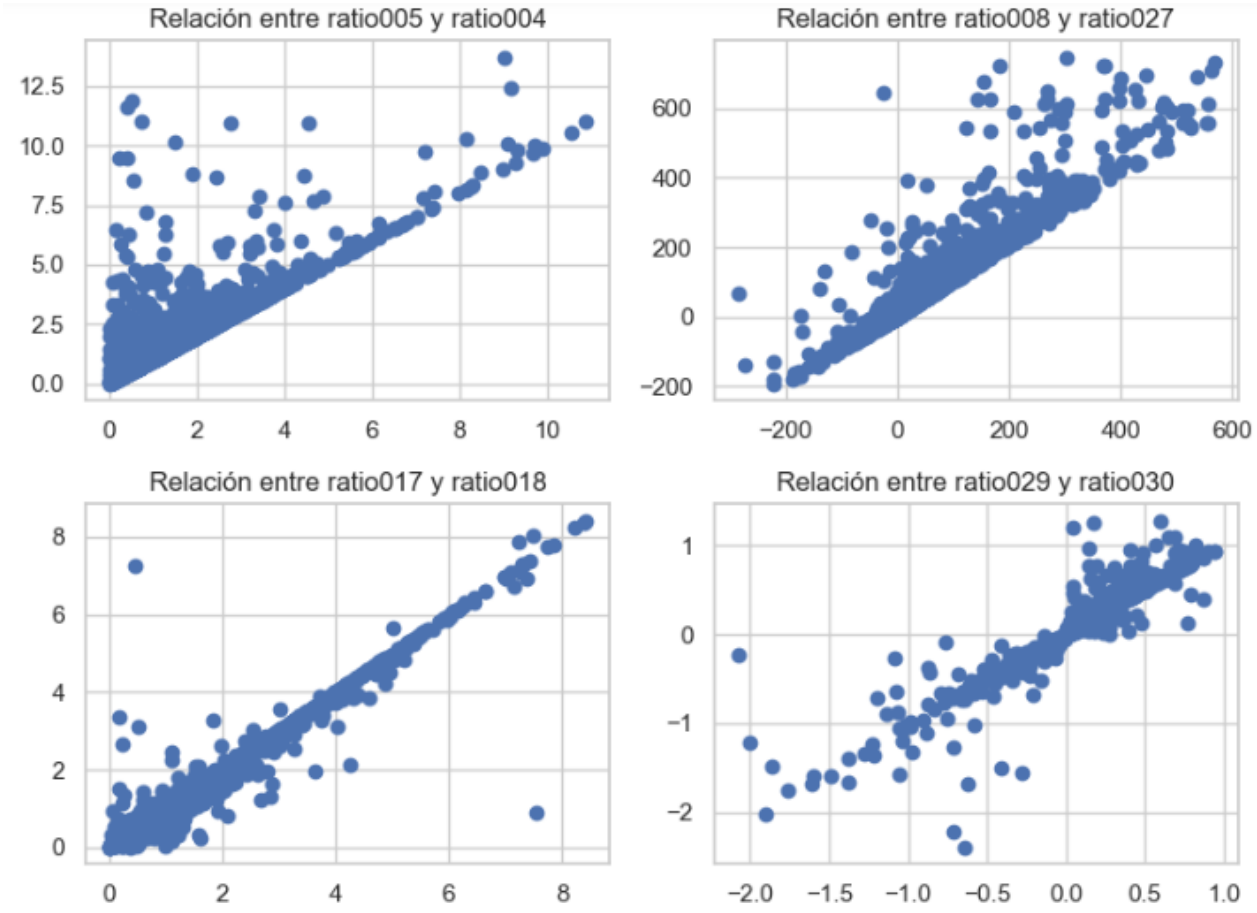
Nos interesa analizar las variables del dataset para obtener información acerca de posible sobre dimensionalidad de la base de datos.



Podemos notar 4 casos en los que el coeficiente de correlación es mayor a 0.8:

- ratio005 y ratio004
- ratio008 y ratio027
- ratio017 y ratio018
- ratio029 y ratio030

Llevamos a cabo un análisis gráfico de estas relaciones



Analizando las variables involucradas en los gráficos podemos notar que la correlación casi perfecta viene dada debido a la construcción de cada ratio. Es decir, en todos los casos las variables comparten numerador o denominador, lo cual genera esta relación lineal.

CONCLUSIONES EXPLORATORY DATA ANALYSIS

- La proporción de empresas en "Default" es del 11% mientras que el restante 89% se encuentran en "No default". Podemos concluir que el dataset es desbalanceado entre clases.
- Del análisis de distribución y test de medias de las variables numéricas del dataset discriminados según status de default o no y se obtuvieron 13 variables cuyas diferencias de medias entre clase son estadísticamente distinta de cero.
- La solvencia de las empresas en "No default" es un 37% mayor en promedio a aquellas en "Default".

- La liquidez de las empresas en "No default" es un 57% mayor en promedio a aquellas en "Default".
- Las empresas en "Default" se apoyan en sus acreedores para generar su resultado operativo en un 110% más en promedio que aquellas empresas en "No default".
- Las empresas en "Default" tienen un 44% más en promedio de ventas a crédito que aquellas empresas en "No default".
- No existen correlaciones relevantes para analizar.

ESTIMACION DE MODELOS

Antes de entrenar los algoritmos y obtener los modelos es importante destacar que el inconveniente que presenta la base de datos con respecto al desbalance de clases puede afectar al rendimiento de los modelos mediante un bajo recall debido a una cantidad relativamente grande de falsos negativos. En el problema particular de este trabajo es muy importante obtener modelos con la menor proporción de falsos negativos posible, ya que es muy costoso para los prestatarios de plataformas P2P otorgar préstamos que son luego defaulteados. En este sentido cada falso negativo "es mas costoso" que un falso positivo (oportunidad de préstamo perdida). Debido al desbalance de clases surgen dos cuestiones:

- Los algoritmos pueden no obtener suficientes datos de la clase minoritaria y por lo tanto resultarles difícil aprender a identificar a dicha clase (en este caso la clase positiva o 1).

- El accuracy del modelo puede no ser una métrica conveniente de utilizar al momento de decidirse entre modelos ya que cuando existen clases desbalanceadas puede haber un accuracy "bueno" (mayor al 90%) y el algoritmo no ser capaz de identificar ningún dato de la clase minoritaria. Es más, si al optimizar los hiperparámetros se indica al accuracy como métrica a maximizar el resultado "óptimo" en este caso podría ser un algoritmo cuyas predicciones sean siempre la clase mayoritaria (0 en nuestro caso), ya que el castigo que recibirá el accuracy por los falsos negativos será muy bajo en relación con el "premio" de identificar todos los valores de la clase mayoritaria.

En base a lo anterior expuesto se decide optimizar los hiperparametros de los algoritmos a entrenar con métricas distintas al simple accuracy. En particular, se utilizará el "balanced accuracy" (el cual tiene en cuenta el desbalance de clase) y el área bajo la curva ROC.

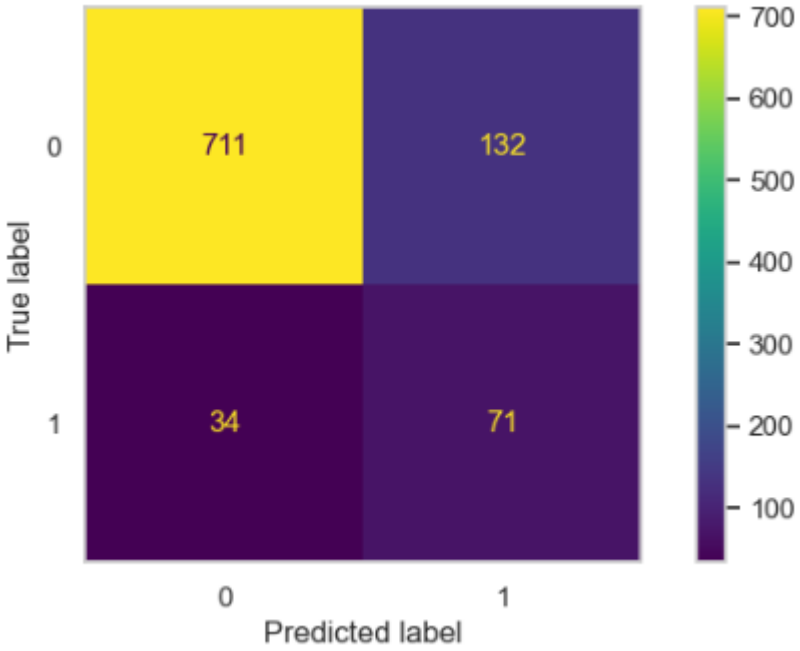
ALGORITMO 1: REGRESION LOGISTICA

La regresión logística es un modelo estadístico utilizado para predecir la probabilidad de un evento binario. Este algoritmo utiliza una función logística para transformar la variable de entrada en una probabilidad. Para nuestro problema de estudio la clasificación que realizará será si el creditó caerá en default (clasificado como 1) o no (clasificado como 0).

Luego de obtener los hiperparametros óptimos para las métricas “balanced_accuracy” y “roc_auc” mediante la funcion RandomizedSearchCV de sklearn y obtener los modelos 1 y 2 respectivamente entrenados con los datos del set de entrenamiento surgen los siguientes resultados de validación:

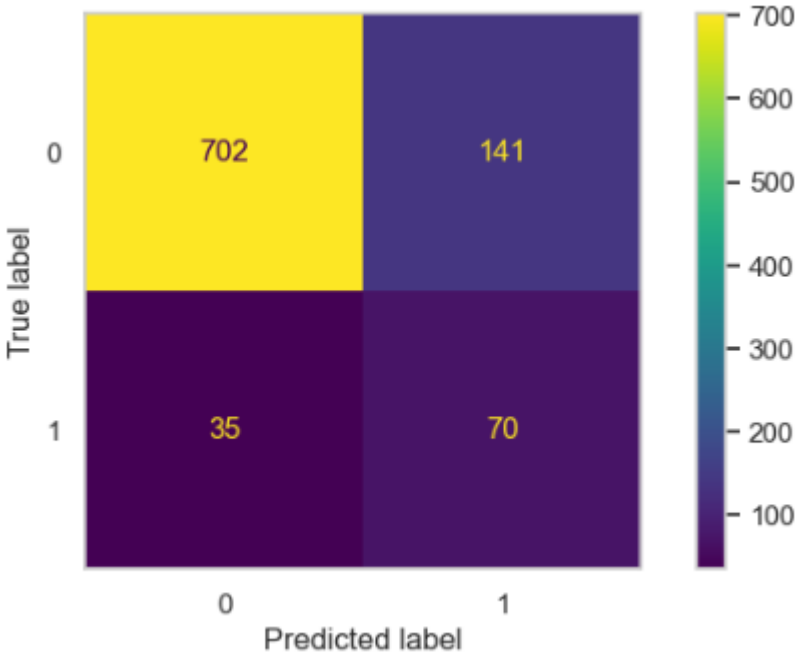
MATRIZ DE CONFUSIÓN Y METRICAS

Modelo 1 (BALANCED_ACCURACY)



	precision	recall	f1-score	support
0	0.95	0.84	0.90	843
1	0.35	0.68	0.46	105
accuracy			0.82	948
macro avg	0.65	0.76	0.68	948
weighted avg	0.89	0.82	0.85	948

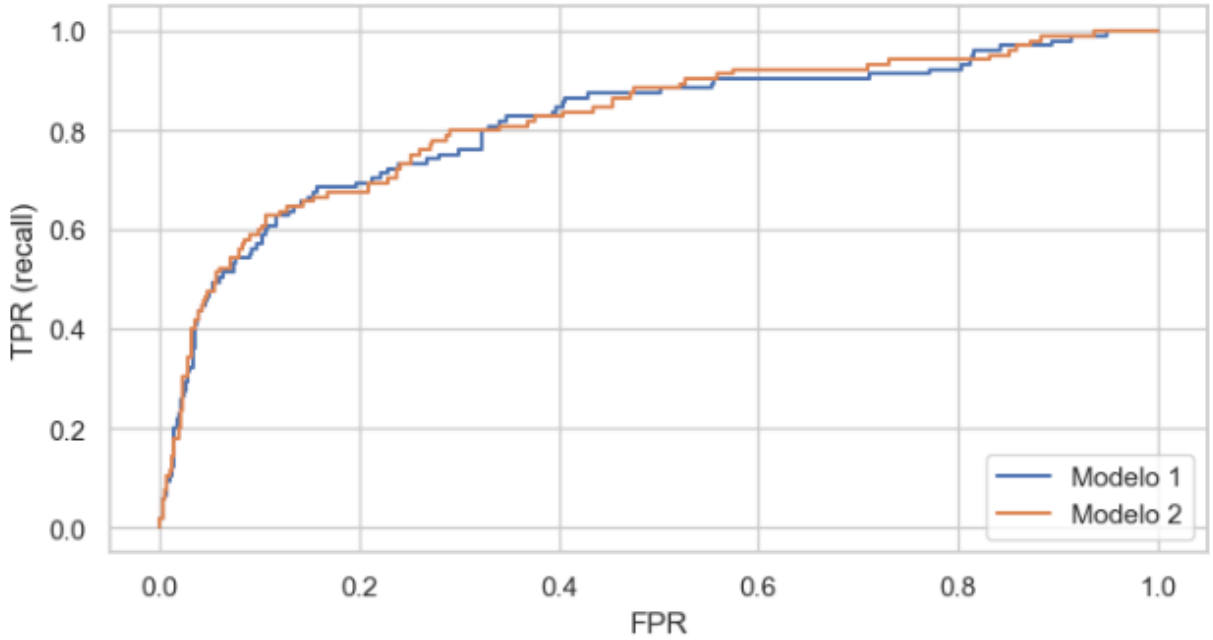
Modelo 2 (ROC_AUC)



	precision	recall	f1-score	support
0	0.95	0.83	0.89	843
1	0.33	0.67	0.44	105
accuracy			0.81	948
macro avg	0.64	0.75	0.67	948
weighted avg	0.88	0.81	0.84	948

El modelo 1 presenta un leve mejor accuracy global (82% vs 83%) asi como un leve mejor recall (68% vs 67%). Esto nos indicaría que el modelo 1 presenta mejor desempeño predictive. Sin embargo, resta analizar la curva ROC y su área ya que es relevante en este caso debido al desbalance de clases.

CURVA ROC MODELO 1 VS MODELO 2



- AUC Modelo 1: 0.814
- AUC Modelo 2: 0.820

La curva ROC muestra cómo varía la tasa de verdaderos positivos (TPR) en función de la tasa de falsos positivos (FPR) a medida que se ajusta el umbral de clasificación del modelo. En la curva ROC, el eje y representa la TPR y el eje x representa la FPR. En este caso, a pesar de que la matriz de confusion del modelo 1 indica mejor desempeño predictive, la curva ROC nos muestra que para distinto umbral de clasificación del modelo (por default es 0,5 en el caso de Regresión Logística) Podemos obtener mejor performance y por lo tanto el modelo 2 es superador.

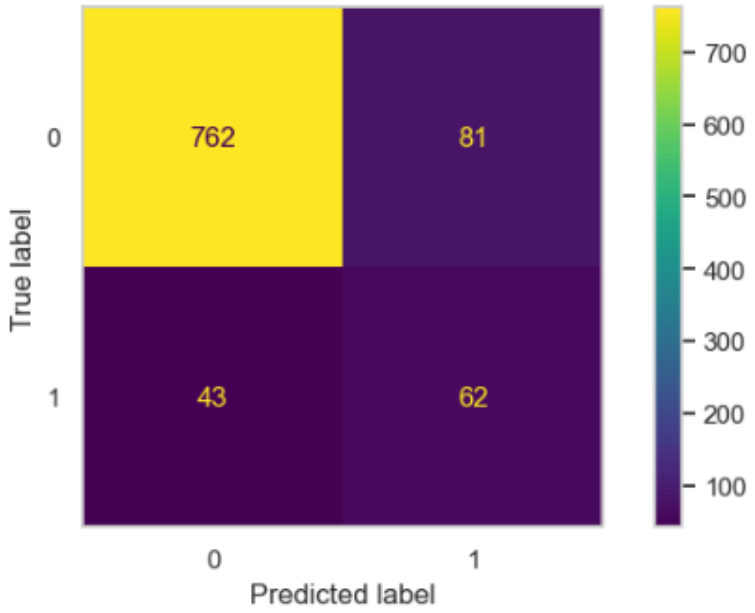
ALGORITMO 2: SUPPORT VECTOR MACHINE

El algoritmo de SVM busca encontrar un hiperplano que separe los datos de diferentes clases de manera óptima mediante el uso de vectores de soporte y maximizando el margen.

Al igual que en el caso anterior buscamos los hiperparametros óptimos para las métricas “balanced_accuracy” y “roc_auc” mediante la funcion RandomizedSearchCV de sklearn y obtener los modelos 1 y 2. Obtenemos los siguientes resultados de validación

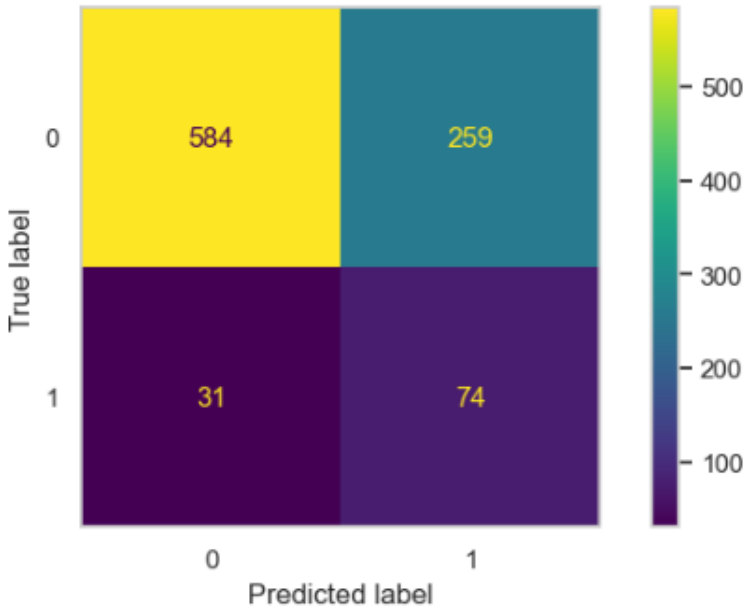
MATRIZ DE CONFUSIÓN Y METRICAS

Modelo 1 (BALANCED_ACCURACY)



	precision	recall	f1-score	support
0	0.95	0.90	0.92	843
1	0.43	0.59	0.50	105
accuracy			0.87	948
macro avg	0.69	0.75	0.71	948
weighted avg	0.89	0.87	0.88	948

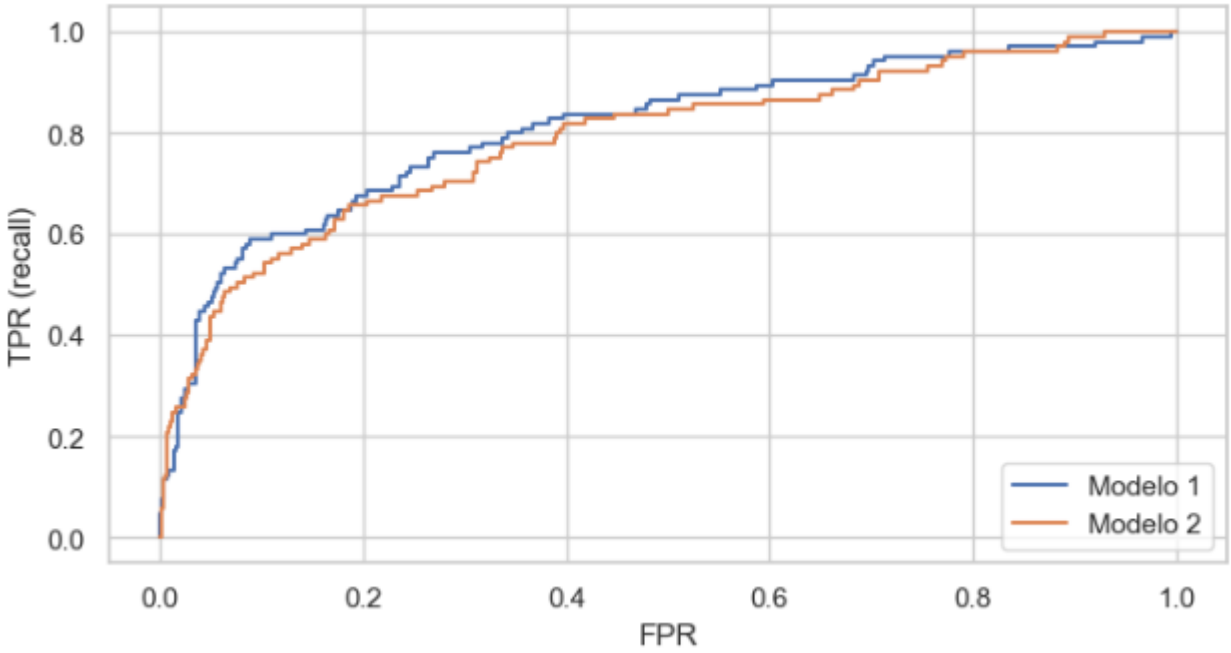
Modelo 2 (ROC_AUC)



	precision	recall	f1-score	support
0	0.95	0.69	0.80	843
1	0.22	0.70	0.34	105
accuracy			0.69	948
macro avg	0.59	0.70	0.57	948
weighted avg	0.87	0.69	0.75	948

Para el algortimo SVC el modelo 2 presenta un peor accuracy global (69% vs 87%) pero un mejor recall (70% vs 59%). Con solo analizar las matrices de confusion no podriamos llegar a una conclusion deifinitiva, ya que a pesar de que el modelo 1 presenta mejor accuracy, el recall es una métrica muy relevante para nuestro problema.

CURVA ROC MODELO 1 VS MODELO 2



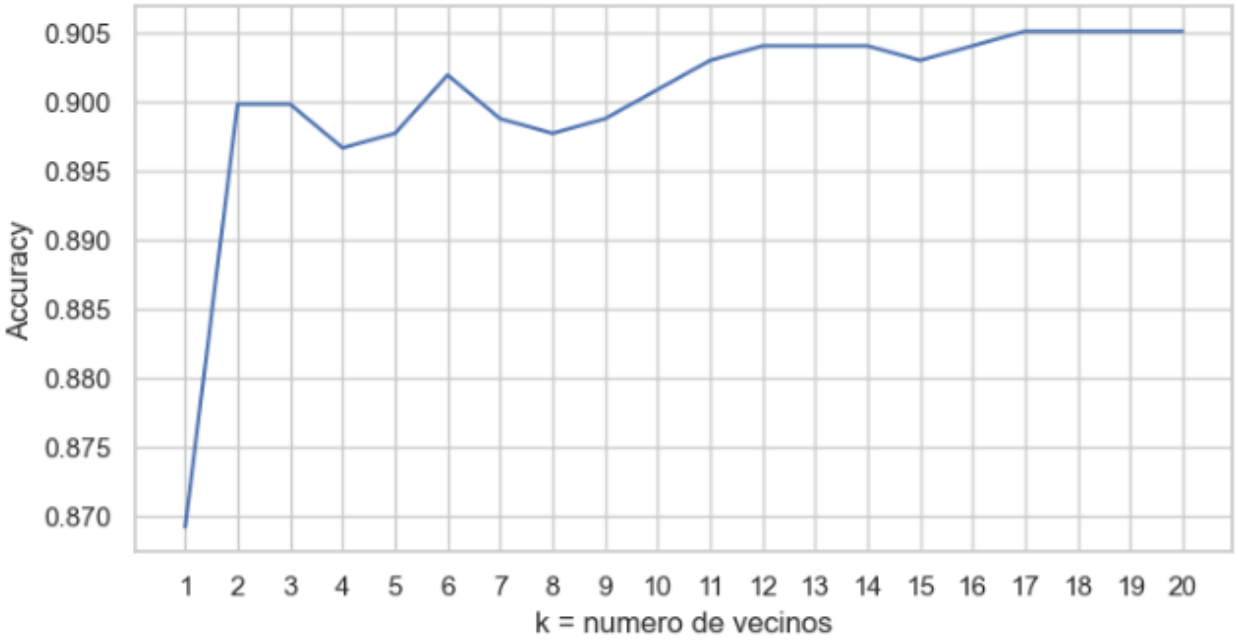
- AUC Modelo 1: 0.809
- AUC Modelo 2: 0.789

En este caso la curva ROC nos muestra que para distinto umbral de clasificación del modelo 1 podemos obtener un mejor TPR por cada FPR, por lo cual el desempeño predictive del modelo 1 es superador.

ALGORITMO 3: K-NEAREST-NEIGHBORS

El algoritmo K-Nearest Neighbors (KNN) es un algoritmo de aprendizaje supervisado utilizado para la clasificación y regresión de datos. Su objetivo es asignar una etiqueta o valor numérico a un punto de datos desconocido basado en los puntos de datos conocidos más cercanos a él.

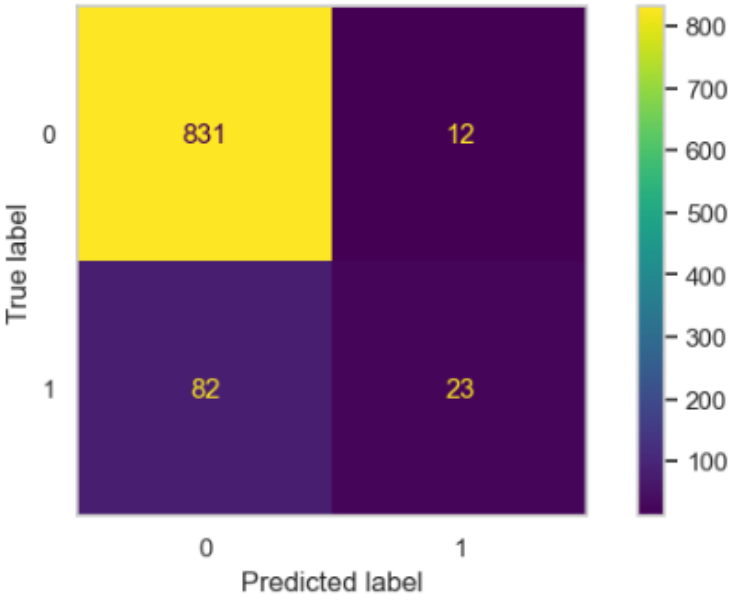
Llevamos a cabo el mismo análisis realizado para los algoritmos anteriores. Sin embargo, uno de los hiperparametros a optimizar es la cantidad vecinos cercanos para clasificar los datos (k), y como no temenos un rango conocido para probar con el RandomizedSearchCV vamos a iterar para distintos valores de k observando el accuracy para cada uno de ellos y asi definir un rango. Se obtiene lo siguiente:



Podemos observar que luego de k = 1 el accuracy del modelo crece y se mantiene en el intervalo (0.895 - 0.905) y a partir de k = 17 el valor parece mantenerse estable. Por lo tanto el intervalo para k sera [2,17].

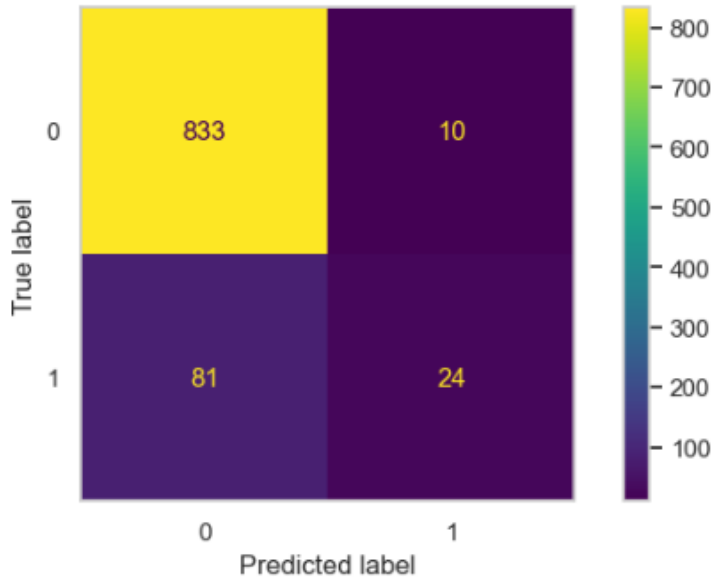
MATRIZ DE CONFUSIÓN Y METRICAS

Modelo 1 (BALANCED_ACCURACY)



	precision	recall	f1-score	support
0	0.91	0.99	0.95	843
1	0.66	0.22	0.33	105
accuracy			0.90	948
macro avg	0.78	0.60	0.64	948
weighted avg	0.88	0.90	0.88	948

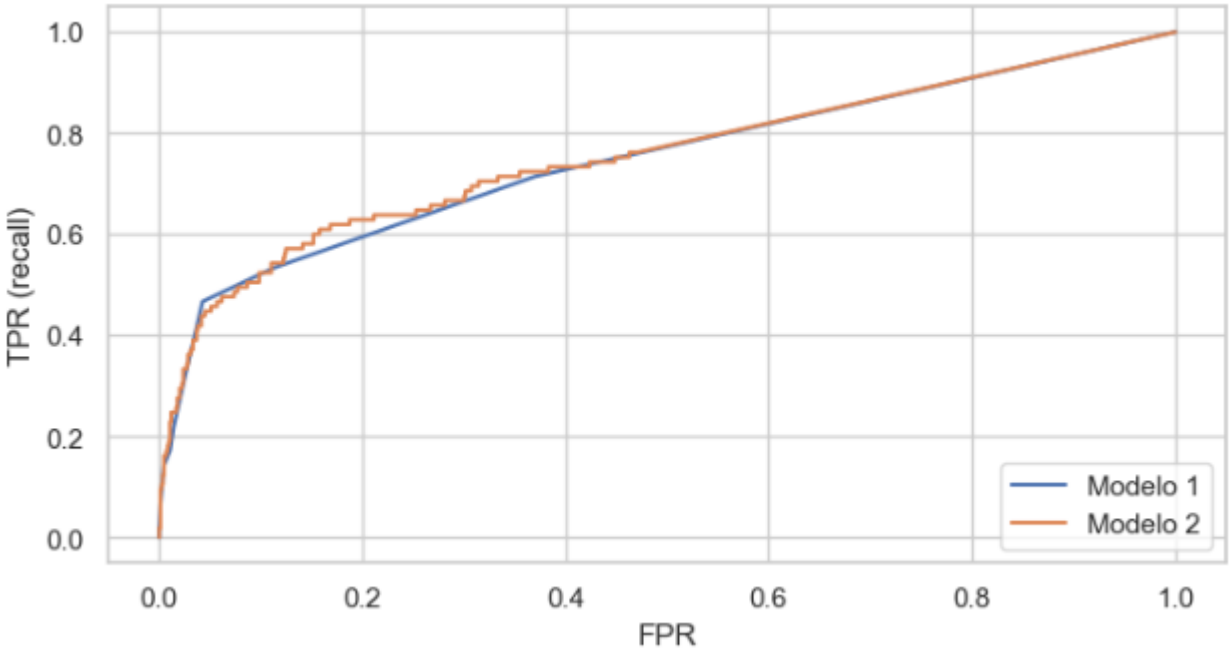
Modelo 2 (ROC_AUC)



	precision	recall	f1-score	support
0	0.91	0.99	0.95	843
1	0.71	0.23	0.35	105
accuracy			0.90	948
macro avg	0.81	0.61	0.65	948
weighted avg	0.89	0.90	0.88	948

Para el algortimo KNN ambos modelos presentan el mismo accuracy (90%) aunque el modelo 2 presenta un recall levemente mayor (23% vs 22%). Ambos modelos son relativamente malos en clasificar a la clase minoritaria.

CURVA ROC MODELO 1 VS MODELO 2



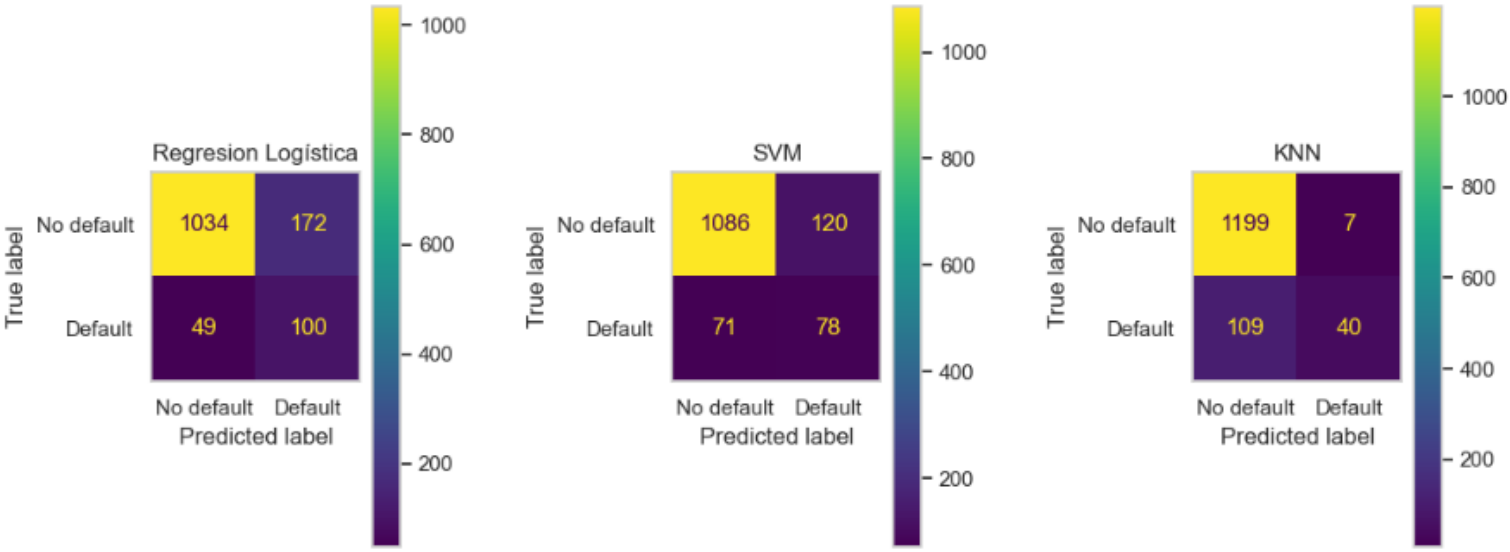
- AUC Modelo 1: 0.747
- AUC Modelo 2: 0.753

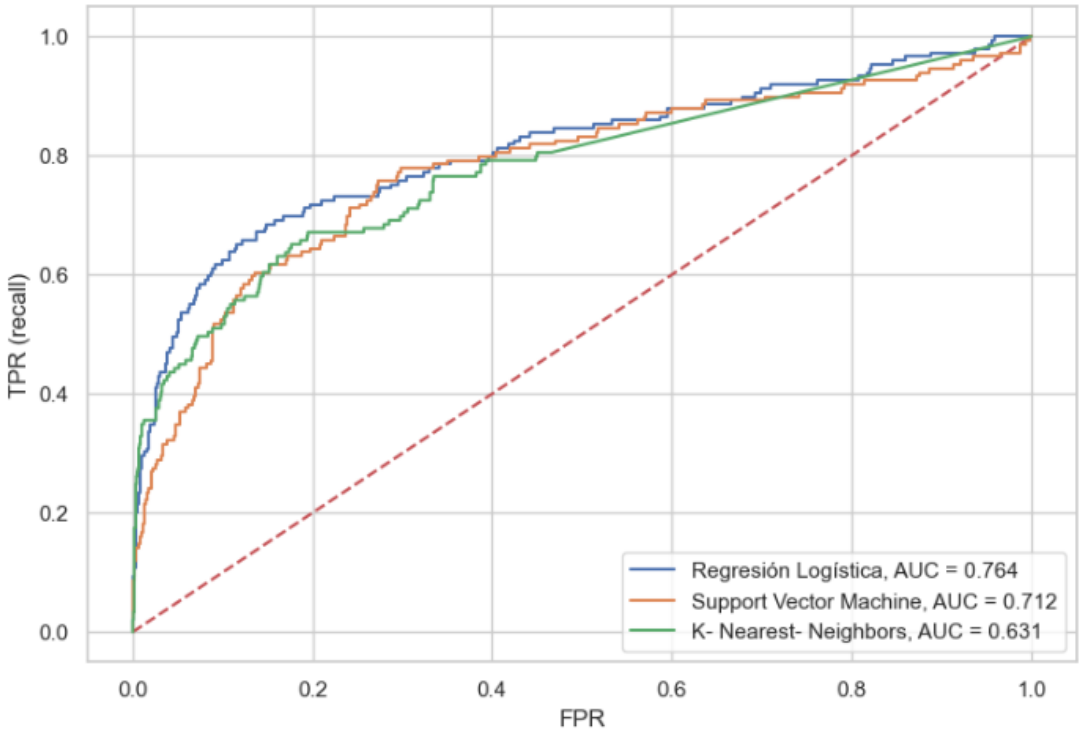
En este caso la curva ROC nos muestra que para distinto umbral de clasificación del modelo 2 podemos obtener un mejor TPR por cada FPR, por lo cual el desempeño predictive del modelo 2 es superador.

EVALUACION Y ELECCION DE MODELO

En esta sección vamos a evaluar con el set de testeo a los tres algoritmos optimizados en base al set de validación. Utilizaremos distintas métricas para la elección del mejor modelo de clasificación.

MATRIZ DE CONFUSIÓN, ROC CURVE Y METRICAS





	Accuracy	Precision	F1 Score	Recall	AUC ROC
Regresión Logística	0.836900	0.367647	0.475059	0.671141	0.764260
SVM	0.859041	0.393939	0.449568	0.523490	0.711994
KNN	0.914391	0.851064	0.408163	0.268456	0.631326

En base al cuadro de métricas podemos notar rapidamente el trade-off que existe (en nuestro caso con clases desbalanceadas) entre accuracy y recall. En este sentido, si nos guiamos por el valor de accuracy la decisión seria quedarnos con el algoritmo de KNN con un accuracy del 91% y un recall del 26%, lo cual implicaría tener un modelo que predice muy bien a la clase mayoritaria (0) pero con grandes dificultades para predecir a la minoritaria (1). Como se mencionó antes, en el negocio analizado el costo de cada uno de los falsos negativos es mayor que el de los falsos positivos por lo que lo óptimo para el caso es prestar mayor atencion al recall.

Del resto de algoritmos, vemos que el algoritmo de Regresión Logística es el que mayor recall presenta (67%) mientras que el accuracy del modelo cae al 83% y la precision es del 36%. Al observar las dos métricas que tienen en cuenta los distintos trafe-offs a los que hay que enfrentarnos: f1-score(precision-recall) y AUC ROC (recall-falsos positivos), notamos que los mayores valores los alcanza para el algoritmo de Regresión Logística (47% y 67% respectivamente).

CONCLUSIÓN

En base al análisis expuesto podemos finalmente elegir al Algoritmo de Regresión Logistica cuyos hiperparámetros optimizan el área bajo la curva ROC como el mejor algoritmo de clasificador en el contexto de evaluación del riesgo crediticio de empresas PyMES que participan en plataformas P2P.