Comprensión del Negocio y Problema

Se obtiene información sobre desembolsos de crédito realizados por un banco que para efectos prácticos llamaremos "Banco Amarillo". Con el objetivo de poder brindar productos bancarios a sus clientes de manera responsable, el mismo desea medir el riesgo para así, controlar la colocación de productos brindados a los clientes de un segmento "objetivo".

Con un conjunto de datos del 2018 que contiene la fecha del desembolso realizado, clientes anónimos en identidad, sin datos demográficos, pero con información relacionada a su comportamiento crediticio y estado "default" o de incumplimiento en pago de obligaciones. El banco desea crear un modelo que pronostique esta probabilidad de incumplimiento en clientes de un conjunto de datos con clientes y desembolsos del 2019-02, y agruparlos de acuerdo con su riesgo crediticio. Y así realizar estrategias accionables y políticas diferenciadas por grupos de riesgo.

Problema: "Se desconoce la probabilidad de pago (riesgo crediticio) de clientes para que el banco pueda ofrecerles productos bancarios de manera segura y segmentada".

Para brindarle a Banco Amarillo visibilidad, conocimiento y una correcta agrupación de su cartera de clientes, se desea crear un modelo para calcular la probabilidad de que un cliente, en los próximos meses, llegue a una altura de mora determinada. Con este objetivo se pondrán a prueba varios modelos de clasificación, utilizando algoritmos de aprendizaje supervisado, que compararemos con métricas de desempeño adecuadas a su objetivo. Y posteriormente, agrupar estos clientes en una cantidad de grupos que permita segmentar la oferta de servicios bancarios.

Comprensión y Preparación de los datos

Se reciben tres archivos del Banco Amarillo:

- Base Train con 28,276 clientes y desembolsos del 2018. Será utilizado para entrenar un modelo predictivo.
- Base Validación con 2,068 clientes y desembolsos del 1er mes del 2019. Será utilizado para poner a prueba el modelo realizado, a través de métricas de rendimiento.
- Base Prueba con 5,889 clientes y desembolsos del 2do mes del 2019. Estos son los clientes descritos en la descripción del problema, para los cuales se desea

pronosticar cuál será su tasa de incumplimiento y a qué grupo de riesgo pertenecen.

Los conjuntos de datos para entrenamiento y validación contienen la variable dependiente, default, con valores:

- “Default”: 1, cliente que incumplió los pagos.
- “No Default”: 0, cliente que cumplió sus pagos.

Cada uno de los conjuntos de datos contiene 27 variables independientes, de distintos tipos: decimal (25), entero (01), texto (01). Estos han sido investigados, adicional al uso del diccionario y planteo lo siguiente que puede tomarse a manera de supuesto o presunción:

- Portafolio Telcos: Productos de telecomunicaciones.
- Portafolio Rotativo: Productos que cuentan con una naturaliza de crédito rotativa, es decir que, a medida que se libera el saldo se renueva el cupo (ejemplo: tarjeta de crédito).
- Portafolio Total: cualquier portafolio de productos.
- Producto Vencido: no pagado a tiempo.
- Producto rotativo cerrado: se asume que el cliente ha pagado la totalidad de su saldo, antes de haber terminado/eliminado/desertado.
- PSE: Sistema de Pagos Electrónicos

Limpieza y Duplicados

Al momento de iniciar el procesamiento y limpieza de datos, se encuentra que no hay datos faltantes, sin embargo, hay 2,277 que se encuentran duplicados en el set de entrenamiento. Esto se debe a que el conjunto de datos no tiene un registro por cliente, sino que tiene un registro de los desembolsos realizados por cliente. Asimismo, se identifica que hay 2,636 clientes duplicados (09%) al unir el set de entrenamiento con el set de validación, es decir que el set de validación tiene 359 transacciones de clientes que se encuentran en el set de entrenamiento. Es importante tratar los clientes duplicados por transacción, porque incluirlos en el modelo podría sesgar el resultado por la conducta generalizada en cuanto al cumplimiento de obligaciones de un cliente. De igual forma, investigo a los clientes con más de una transacción en cuanto a la similitud en sus variables y una mejor comprensión de ellas – como resultado, concluyo que vale la pena eliminarlos, debido a que algunas variables se mantienen constantes en todas las transacciones del mismo cliente. Una variable que se mantiene constante correctamente es: Tipo_cliente.

Aunque elimino, conservaré de las transacciones duplicadas, aquella en que esté default = 1, así favorezco la categoría que es más escaza en el conjunto de datos.

En la búsqueda de variables con presencia de ceros, encuentro estas dos con mayor cantidad:

- sd_cantidad_retiros_12m (AH) tiene 9645 (41.3%) ceros.
- sd_monto_recarga_12m (TO) tiene 9588 (41.1%) ceros.

Dado que ambas variables me indican que corresponde a la desviación estándar en una acción del cliente, el cero no me indica ausencia de validez en la observación, sino que los clientes han tenido consistencia en su ejecución.

Outliers

Durante mi análisis de valores atípicos, grafico algunas variables de interés para tener visibilidad sobre su dispersión y qué tanto ruido tiene el dataset de entrenamiento, a rasgos generales.

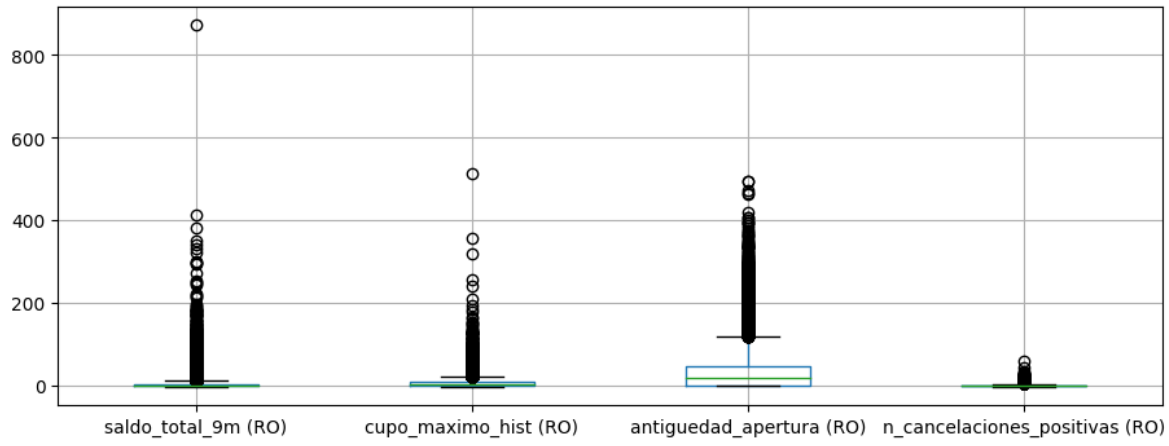

Imagen_01: Variables de Interés

Como se puede observar en la Imagen 01, las variables que fueron elegidas por su interpretabilidad representadas en diagramas de caja muestran una cantidad significativa de datos que se encuentran alejados de los bigotes, representando la variabilidad de la muestra. Esto debe ser atendido antes de la construcción del modelo, sin afectar negativamente la cantidad de datos que se tienen para entrenamiento.

Una variable que me interesa tener correctamente es: saldo_total_9m (RO) / CO01END051RO; dado que me podría indicar qué tan saludable está la cartera del banco, por cliente. Calculo el rango intercuartílico y su límite superior e inferior. Y para decidir cuánto ruido puedo retirar, utilizo un factor “k” = 1.5. Con este análisis debería

eliminar 2,903 observaciones. Aumento el valor de k a 3, para reducir a 1,730 observaciones a eliminar – todavía elevado (7% del dataset) si considero que solo es una variable de 27.

Optimizo este análisis por medio de una iteración y tomo acción dependiendo del porcentaje del total que representa la cantidad de outliers de la variable:

- **>10% eliminar variable:** 'monto_min_transado_1m (TO)', 'monto_min_transado_2m (TO)'.
- **5-10% considerar eliminar observaciones:** 'sd_monto_recarga_12m (TO)', 'retiro_monto_promedio_12m (AH)', 'saldo_total_9m (RO)'.
- **0-5% revisar:** 'sd_cantidad_retiros_12m (AH)', 'os_cel', 'cupo_promedio_hist (RO)', 'ultima_cancelacion_positiva_m (CC)', '%_antiguedad_>48m (TO)', 'recencia_apertura (TO)', 'antiguedad_apertura (CC)', 'antiguedad_apertura (RO)', '%_cartera_vencida (CC)', 'saldo_promedio_hist (RO)', 'frecuencia_uso_3m (RO)', 'cupo_maximo_hist (RO)', 'participacion_ctas_ahorro (TO)', '%_cierre_del_total (RO)', 'recencia_apertura (AH)', 'cantidad_cta_ahorro (AH)', 'n_cancelaciones_positivas (RO)', 'ponderacion_reportes_positivo_18m (TO)', 'ponderacion_reportes_positivo_24m (RO)'.

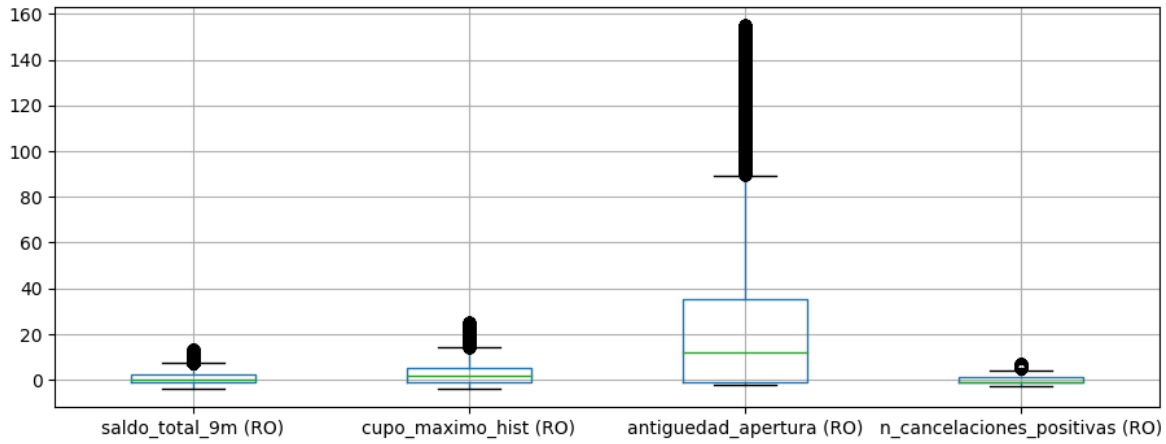

Imagen_02: Variables de Interés con reducción de ruido.

Al observar el eje y de la Imagen 02, se observa que gracias a la eliminación de outliers, ahora la dispersión de las observaciones de variables de interés es menor. Asimismo, ahora es más visible la mediana (línea verde) y se reducen los valores atípicos fuera de bigotes.

Después de eliminar los datos atípicos y duplicados, el tamaño del dataset se redujo en un 17.5% (10% por outliers) y queda con 25 variables explicativas sobre el portafolio de productos: crédito rotativo, telecomunicaciones y ahorro.

Análisis Descriptivo

En esta sección del entregable utilizo herramientas de analítica descriptiva para brindar una mayor comprensión de los clientes del banco que han recibido desembolsos de crédito.

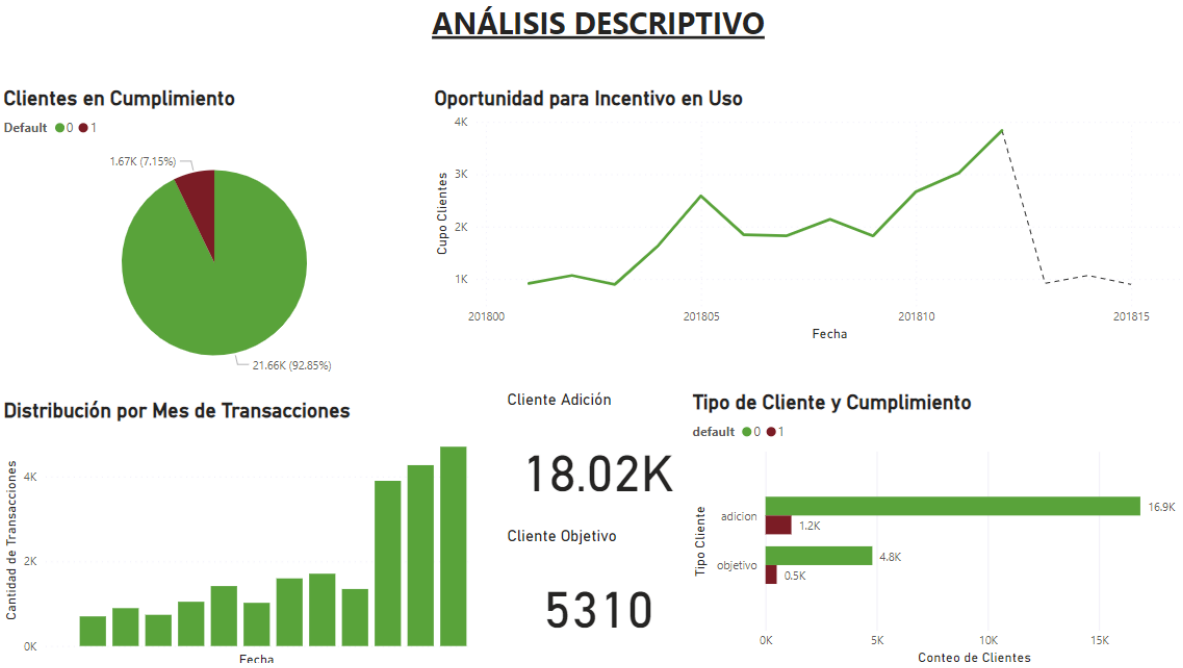

Imagen_03: Análisis Descriptivo

Generales:

- 23,332 clientes únicos.
- Se reduce el set de entrenamiento en un 17.5% entre duplicados y outliers.
- 25 variables explicativas sobre portafolio de productos: crédito rotativo, telecomunicaciones y ahorro.
- Mas de la mitad de los registros pertenece a los últimos 03 meses del año 2018.
- 77% son del tipo de cliente "Adición".
- 23% son del tipo de cliente "Objetivo".
- 93% de los clientes se encuentran en cumplimiento con el pago de sus obligaciones.

Hay un desbalance importante en la clase "No Default". Esto lo vamos a considerar durante la fase de modelamiento, ya que el modelo va a buscar reducir el error pronosticando cercano a 0.

Se encuentra poca interpretabilidad de algunas variables debido a sus valores. Por ello se realiza una revisión de la media y desviación estándar, en búsqueda de valores que indiquen que las variables vengan de otras bases de datos en que fueron sometidas a algún método de normalización- no es el caso.

Debido a valores polarizados sin límites claros, valores negativos sin sentido (%), amplia variabilidad, valores monetarios muy bajos, ausencia de un patrón de escalas categóricas; se infiere que pudo haber un error durante la recolección o extracción de datos, y debería ser investigado.

Correlación

Dicho lo anterior, procedo a investigar la correlación entre sus variables y conocer si su fuerza y dirección aún puede interpretarse en sus datos. Ya que la mayoría de las variables son continuas, utilizo la correlación de Pearson para este ejercicio. Anoto observaciones sobre este análisis, también con el objetivo de prevenir colinealidad en variables del modelo.

Nota: Considerando que mi variable dependiente, default, es dicotómica – no la considero en esta sección por no ser una metodología apropiada. Evaluaré la misma en la sección: Selección de Variables.

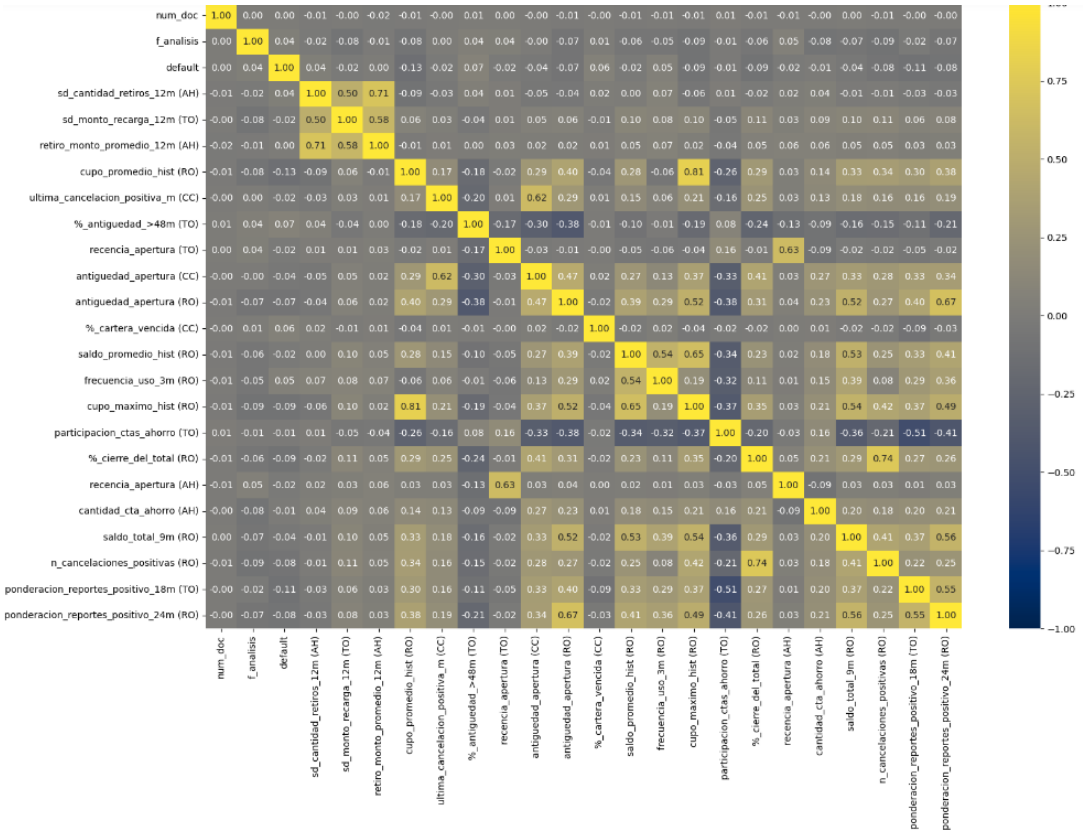

Imagen_04: Matriz de Correlación

Portafolio Ahorro:

- **sd_cantidad_retiros_12m (AH) tiene una correlación positiva fuerte con retiro_monto_promedio_12m (AH):** puede darse a la naturaleza del cálculo de ambas. Se elimina para evitar colinealidad y por p-value elevado.

Portafolio Rotativo:

- **cupo_promedio_hist (RO) tiene una correlación positiva con antiguedad_apertura (RO)**: lo que me puede indicar que los clientes más antiguos pueden llegar a obtener límites de crédito más altos, o tienen un mejor comportamiento de pago de obligaciones.
- **cupo_promedio_hist (RO) tiene una correlación positiva fuerte con cupo_maximo_hist (RO):** los clientes que mantienen un promedio alto para consumo de crédito (sea porque pagan o usan poco), tienen límites de crédito mayor. Contiene valores atípicos (400). Podría eliminar cupo_promedio_hist (RO).
- **n_cancelaciones_positivas (RO) tiene una fuerte correlación positiva con %_cierre_del_total (RO):** un cliente que realiza más pagos positivos (dentro del

plazo y del monto), va a tener un % de cierre del total mayor. Aunque hay riesgo de colinealidad, mantengo por su baja variabilidad en el conjunto de datos.

- **frecuencia_uso_3m (RO) tiene cierta correlación positiva con antiguedad_apertura (RO):** podría indicar que los clientes más antiguos utilizan más sus productos rotativos, sin embargo, la correlación no es lo suficientemente fuerte para inferirlo.
- **ponderacion_reportes_positivo_18m (TO) tiene una correlación positiva con antiguedad_apertura (RO):** los clientes que tuvieron mayores reportes positivos en 18 meses suelen ser más antiguos.
- **ponderacion_reportes_positivo_24m (RO) tiene una correlación positiva con antiguedad_apertura (RO):** Refuerza el punto anterior, en esta, únicamente para portafolio rotativo.
- **ponderacion_reportes_positivo_24m (RO) tiene una correlación positiva con ponderacion_reportes_positivo_18m (TO):** puede explicarse por una proporción considerable de productos rotativos.

Portafolio Telcos:

- **ultima_cancelacion_positiva_m (CC) tiene una correlación positiva con antiguedad_apertura (CC):** esto me puede indicar que los clientes Telcos más antiguos, han pagado positivamente en el pago más reciente.

Portafolio Total:

- **recencia_apertura (TO) tiene una correlación positiva con recencia_apertura (AH**): podría considerar que los productos de ahorro son mayores en proporción que los otros productos, por ello el producto total aumenta junto con el de ahorro.

Planteamiento de Hipótesis

En el análisis y experimentación con el dataset del 2018, la variable default con dos grupos: default y no default, es utilizada como variable de interés para pronosticar y agrupar, por lo que previo a la elaboración del modelo, es imperativo investigar su origen a partir de variables explicativas. Por ello se evaluará la correlación con las variables independientes y así concluir si son estadísticamente significativas explicando a la variable dependiente, para ser utilizadas en el modelo predictivo.

Hipótesis Nula

H_0: No existe una relación significativa entre la variable dependiente, default y las variables explicativas que se encuentran en el dataset. Es decir que las variables

independientes o explicativas no podrían utilizarse para explicar la probabilidad de cumplimiento en obligaciones de pago, por ser igual a cero o muy cercano al cero.

Modelo de Clasificación

Selección de Variables

Primero realizo una selección de variables basada en interpretabilidad y funcionalidad, sin afectar la capacidad de predicción del modelo. Posteriormente analizo las variables en orden de P-Value, con relación a la hipótesis nula:

- 'retiro_monto_promedio_12m (AH)': **retirar**.
- 'monto_min_transado_1m (TO)': **excluida por su cantidad de outliers y por irrelevancia**.
- 'cantidad_cta_ahorro (AH)': **retirar**.
- 'participacion_ctas_ahorro (TO)': **retirar**.
- 'monto_min_transado_2m (TO)': **excluida por cantidad de outliers y por irrelevancia**.
- os_cel: es estadísticamente significativa para la variable dependiente, con un p-value muy bajo. Es poco probable que sea azar, aunque podría deberse a una correlación espuria. Por el valor que añade en poder predictivo, permanecerá en el modelo, por lo menos hasta la fase de validación.
- Las variables de identificación de cliente, fecha de transacción y tipo de cliente (categórica) no son incluidas en este análisis.

Hipótesis Alternativa

H1: Existe evidencia suficiente para rechazar la hipótesis nula, ya que se encontraron 19 variables independientes con elevada relación con la variable dependiente, default.

	correlation	p_value
cupo_promedio_hist (RO)	-0.127129	1.169011e-84
ponderacion_reportes_positivo_18m (TO)	-0.105114	2.561985e-58
cupo_maximo_hist (RO)	-0.090059	3.197436e-43
os_cel	-0.089894	4.543155e-43
%_cierre_del_total (RO)	-0.088251	1.444400e-41
ponderacion_reportes_positivo_24m (RO)	-0.080867	3.707079e-35
n_cancelaciones_positivas (RO)	-0.077387	2.486774e-32
%_antiguedad_>48m (TO)	0.072068	2.992197e-28
antiguedad_apertura (RO)	-0.071849	4.341148e-28
%_cartera_vencida (CC)	0.057551	1.397799e-18
frecuencia_uso_3m (RO)	0.051581	3.176293e-15
saldo_total_9m (RO)	-0.038553	3.843044e-09
sd_cantidad_retiros_12m (AH)	0.037306	1.197081e-08
antiguedad_apertura (CC)	-0.036157	3.304650e-08
recencia_apertura (AH)	-0.022538	5.756101e-04
recencia_apertura (TO)	-0.021875	8.329245e-04
sd_monto_recarga_12m (TO)	-0.021456	1.046807e-03
ultima_cancelacion_positiva_m (CC)	-0.021382	1.090075e-03
saldo_promedio_hist (RO)	-0.018819	4.043556e-03

Variables significativas: 19

Imagen_05: Evidencia estadística

Modelo de Clasificación

Regresión Logística

Utilizo un algoritmo de clasificación para entrenar el conjunto de datos train. Divido el conjunto en dos grupos: 'default' y 'no default'. Primero observo las métricas de evaluación del modelo aplicado al mismo y posteriormente al conjunto de validación.

Para la evaluación del conjunto de datos inicial, utilizo un umbral en que menos 0.3 es 'no default' (T1-T7), y mayor o igual a 0.3 es 'default' (T8). Este umbral me permite evaluar la capacidad del modelo en predecir dos clases, únicamente. Un umbral menor abarca una mayor cantidad de predicciones de la clase no predominante: 'default'. Desde su defecto (0.5) a este nuevo umbral, la precisión (+) baja de 0.44 a 0.34, y el recall (+) sube de 0.005 a 0.03. Me interesa subir precisión (+) y recall (+), porque quiero saber cuántas de las predicciones positivas fueron correctas (precisión) y de los valores positivos, cuántos pude predecir correctamente (recall). Por el riesgo que conlleva, deseo disminuir Falsos-Positivos.

- Falsos-Positivos: Me equivoco dando productos riesgosos a clientes con mayor riesgo del predicho - pérdida por desacato y costos administrativos de cobro.

- Falsos-Negativos: Dejo de darle crédito o le asigno un grupo de riesgo mayor del debido, limitándole la cartera de productos - costo de oportunidad.

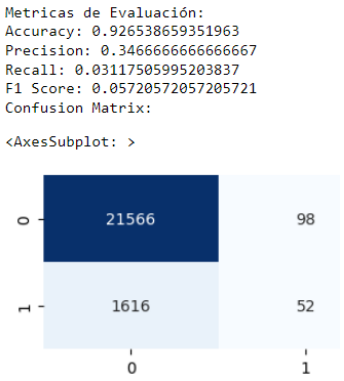

Imagen_06: Métricas de Evaluación T=0.3

Al aplicar este modelo al conjunto de validación, obtengo los siguientes resultados.

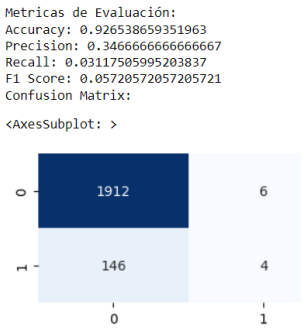

Imagen_07: Métricas de Evaluación V=0.3

Evaluación para Asignación de Grupos

Asigno los grupos de riesgo de acuerdo con el condicional descrito para así segmentar y ofrecer productos por grupos. Para poder validar la capacidad del modelo en colocar en el grupo correcto, utilizaré los clientes duplicados entre la base de entrenamiento y en la base de validación, 318 clientes.

Con esto logro comparar cuántos de los clientes fueron asignados al mismo grupo cuando se aplicó el mismo modelo al set de entrenamiento y de validación. Con fines gráficos muestro una imagen del resultado. El accuracy por grupo para el modelo utilizando regresión logística tuvo un: **43% de Accuracy**.

```
df_grupos[:10]
```

	num_doc	grupo_train	grupo_validacion	acc
0	104630908229	T6	T6	True
1	10627359005	T6	T6	True
2	111504450365	T7	T6	False
3	115062512307	T4	T4	True
4	119214748986	T4	T5	False
5	122109774016	T4	T2	False
6	124776742865	T8	T8	True
7	130768631561	T5	T4	False
8	131370404644	T3	T3	True
9	133176873743	T5	T5	True

Imagen_08: Evaluación por Grupos

KNN

Considerando que la naturaleza del algoritmo K-Nearest Neighbors busca asignar clases, y que en un problema de clasificación se basa en la mayoría de la incidencia en sus 'k' vecinos más cercanos. Con esto quiero decir que mi resultado va a ser dicotómico, al igual que la variable dependiente; y que, ya que la asignación de grupos necesita la probabilidad como una variable continua, utilizaré un método se sklearn.

Con "K"=1 encuentro un accuracy y precisión(+) casi perfecto en el set de entrenamiento. Esto me indica sobreajuste y con una cantidad de venios tan baja estamos captando el ruido del dataset. Adicional, sería altamente complejo y poco generalizable al conjunto de prueba.

Con "K"=3 no está tan generalizado , el valor fue elegido utilizando el diagrama de codo. A continuación los resultados en el set de validación:

```
Metricas de Evaluación:
Accuracy: 0.9387107834733414
Precision: 0.7195571955719557
Recall: 0.23381294964028776
F1 Score: 0.35294117647058826
Confusion Matrix:

<AxesSubplot: >
```

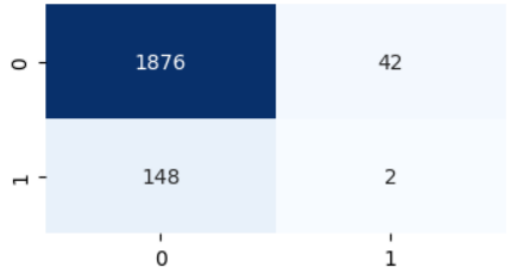

Imagen_09: Métricas de Evaluación V

Evaluación para Asignación de Grupos

El accuracy por grupo para el modelo utilizando KNN tuvo un: 83% de Accuracy**.** Esto me indica que este modelo tiene un accuracy superior que la regresión logística para poder asignar a los clientes en varios grupos. Sin embargo tiene una gran deficiencia, como fue explicado en la introducción al algoritmo, no da valores continuos , sino que por usar los valores cercanos, sus opciones son: 0, 0.33, 0.66, 1.

```
df_grupos_lr[:10]
```

	num_doc	grupo_train_lr	grupo_validacion_lr
0	104630908229	T6	T6
1	10627359005	T6	T6
2	111504450365	T7	T6
3	115062512307	T4	T4
4	119214748986	T4	T5
5	122109774016	T4	T2
6	124776742865	T8	T8
7	130768631561	T5	T4
8	131370404644	T3	T3
9	133176873743	T5	T5

Imagen_10: Evaluación por Grupos

Conclusión

El modelo ganador (LR) requiere seguir siendo entrenado con nuevos datos y optimizado en sus parámetros. Es necesario un “after action report” y considerar ante su uso que ha sido un dataset poco diverso y con mucha variabilidad en sus datos.

Con estos hallazgos se recomienda ofrecer productos menos riesgosos a los clientes en límites más altos, o temporalmente subdividir los grupos de riesgo en mayor cantidad de categorías: T8-T12.

Anexos

Anexo_01: Creación de diccionario

Campo	Rename
CO01ACP011RO	n_cancelaciones_positivas (RO)
CO01ACP017CC	ultima_cancelacion_positiva_m (CC)
CO01END002RO	saldo_promedio_hist (RO)
CO01END010RO	cupo_promedio_hist(RO)
CO01END051RO	saldo_total_9m (RO)
CO01END086RO	frecuencia_uso_3m (RO)
CO01END094RO	cupo_maximo_hist (RO)
CO01EXP001CC	antiguedad_aperterua (CC)
CO01EXP002AH	recencia_apertura (AH)
CO01EXP003RO	antiguedad_apertura (RO)
CO01MOR098RO	ponderacion_reportes_positivo_24m (RO)
CO01NUM002AH	cantidad_cta_ahorro (AH)
CO02END015CC	%_cartera_vencida (CC)
CO02EXP004TO	recencia_apertura (TO)
CO02EXP011TO	%_antiguedad_>48m (TO)
CO02MOR092TO	ponderacion_reportes_positivo_18m (TO)
CO02NUM043RO	%_cierre_del_total (RO)
CO02NUM086AH	participacion_ctas_ahorro (TO)
disp309	os_cel
trx102	monto_min_transado_1m (TO)
trx106	monto_min_transado_2m
trx143	sd_monto_recarga_12m
trx158	retiro_monto_promedio_12m (AH)
trx39	sd_cantidad_retiros_12m
num_doc	num_doc
f_analisis	f_analisis
default	default
tipo_cliente	tipo_cliente