

Proyecto 3 Riesgo Relativo

Hito 1

Sprint Planning 1

¿Qué paso es nuestra principal prioridad en este sprint?

Leer y entender el proyecto para saber con certeza cómo lo voy a abordar.

¿Qué tareas se desprenden de este paso? ¿Manejan alguna dependencia?

- ☒ ~~Leer toda la información del proyecto~~
- ☒ ~~Hacer anotaciones que me puedan ayudar en el futuro~~
- ☐ Comenzar el primer paso que es el **procesado y preparación de la base de datos**.
- ☐ Al menos alcanzar algunos puntos del **análisis exploratorio**.

Planteamiento del proyecto

Tenemos al banco “Super Caja” que ha disminuido las tasas de interés, por lo cual ha generado un notable aumento en la demanda de crédito. Sin embargo, el banco no estaba preparado para ello y ha sobrecargado al equipo de análisis de crédito, el cual se encuentra en un proceso manual e ineficiente.

Para ello se requiere armar un score crediticio a partir de un análisis de datos y evaluación del riesgo relativo que pueda clasificar a los solicitantes en diferentes categorías de riesgo, basadas en su probabilidad de incumplimiento. Esta clasificación permitirá al banco tomar decisiones informadas sobre a quién otorgar crédito, reduciendo así el riesgo de préstamos no reembolsables. Además la integración de la métrica existente de pagos atrasados fortalecerá la capacidad del modelo para identificar riesgos, lo que en última instancia contribuirá a la solidez financiera y la eficiencia operativa del banco.

El **desafío** es diseñar la automatización del proceso de análisis mediante avanzadas técnicas de análisis de datos.

El **objetivo** principal es mejorar la eficiencia y la precisión en la evaluación del riesgo crediticio, permitiendo al banco tomar decisiones informadas sobre la concesión de crédito y reducir el riesgo de préstamos no reembolsables.

Esta propuesta también destaca la integración de una métrica existente de pagos atrasados, fortaleciendo así la capacidad del modelo.

Habilidades: Análisis de datos, clasificación de clientes, matriz de confusión, consultas complejas en BigQuery.

La **matriz de confusión** permitirá evaluar cuán efectivas son tus reglas y cómo se están comportando en la clasificación de clientes, lo que es esencial para la toma de decisiones en el análisis de crédito.

Herramientas y/o plataformas

- Google BigQuery
- Google Colab: Plataforma para trabajar con Python y Notebooks.
- Google Slides
- Google Looker Studio: Herramienta para la creación y edición de dashboards, informe de datos.

Lenguajes

- SQL y Python

Insumos

El conjunto de datos contiene información sobre préstamos concedidos a un grupo de clientes del Banco. Los datos se dividen en 4 tablas.

1. **user_info**: Datos del usuario/cliente.

user_id	Número de identificación del cliente (único para cada cliente)
age	Edad del cliente
sex	Sexo del cliente
last_month_salary	Último salario mensual que el cliente reportó al banco
number_dependents	Número de dependientes

2. **loans_outstanding**: Datos del tipo de préstamo.

loan_id	Número de identificación del préstamo (único para cada préstamo)
user_id	Número de identificación del cliente
loan_type	Tipo de préstamo (real estate = inmobiliario, others = otro)

3. **loans_detail**: Datos del comportamiento de pagos de estos préstamos.

user_id	Número de identificación del cliente
more_90_days_overdue	Número de veces que el cliente estuvo más de 90 días vencido

using_lines_not_secured_personal_assets	Cuánto está utilizando el cliente en relación con su límite de crédito, en líneas que no están garantizadas con bienes personales, como inmuebles y automóviles
number_times_delayed_payment_loan_30_59_days	Número de veces que el cliente se retrasó en el pago de un préstamo (entre 30 y 59 días)
debt_ratio	Relación entre las deudas y el patrimonio del prestatario. $\text{ratio de deuda} = \frac{\text{Deudas}}{\text{Patrimonio}}$
number_times_delayed_payment_loan_60_89_days	Número de veces que el cliente retrasó el pago de un préstamo, (entre 60 y 89 días)

4. **default:** Identificación de clientes ya identificados como morosos.

user_id	Número de identificación del cliente
default_flag	Clasificación de los clientes morosos (1 para clientes que pagan mal, 0 para clientes que pagan bien)

Introducción

El riesgo relativo es una medida estadística, utilizada en epidemiología y en otras áreas para evaluar cuánto más probable es que ocurra un resultado en el grupo expuesto en comparación con el grupo no expuesto.

Problema que enfrentamos: La necesidad de automatizar y optimizar el proceso de análisis crediticio para gestionar eficazmente el riesgo de incumplimiento.

Preguntas clave: ¿Qué variables influyen más en el riesgo de incumplimiento? ¿Cómo se correlacionan estas variables entre sí y con el comportamiento de pago de los clientes?

Riesgo relativo = 1 Sugiere que no hay diferencia entre los dos grupos.

Riesgo relativo > 1 Indica un mayor riesgo en el grupo expuesto.

Riesgo relativo < 1 Indica un menor riesgo en el grupo expuesto.

El análisis en este hito sentará las bases para una evaluación crediticia más precisa y automatizada que, en última instancia, contribuirá a la solidez financiera y operativa del banco.

Procesar y preparar la base de datos

Meta	Herramienta	Fecha	Explicación	Comentarios/Preguntas
Conectar /Importar Datos A Herramientas	Big Query	13/12/2023	Cree un proyecto que se llama lucid-vector-408016 (supongo que me puso ese por que el que elegí estaba repetido), El conjunto de datos se llama super_caja_ds . Finalmente cargue las 4 tablas.	NOTA: Recordar dejar la región y vencimiento para el conjunto de datos que te dan por default. Para crear las tablas importante poner detección automática
Identificar y manejar valores nulos	Big Query	13/12/2023	Tengo 7199 registros nuls en last_month_salary que representa el 20% de mi muestra total. Sin embargo, al unir la tabla con default y solamente mandar llamar los clientes que NO pagan y los que tienen valor nulo en su último salario reportado nos arroja 130 registros los cuales representan un 0.36% de mi muestra total por lo que decidí eliminarlos de mis registros ya que no considero que sea de interés del banco tener clientes que no pagan y no saber cuánto ganan mensualmente.	Porcentaje=(no. registros de interés/Total de registros)x100 Lo guardé en la consulta user-default_JOIN
Identificar y manejar datos fuera del alcance del análisis	Big Query	14/12/2023	Utilice el comando siguiente para obtener los duplicados de los user_id y en el caso de la tabla loans_outstanding utilicé el loan_id	Utilizando SELECT user_id, COUNT (user_id) AS total FROM `lucid-vector-408016.super_caja_ds.default` GROUP BY user_id HAVING COUNT (user_id) > 1;
Identificar y manejar datos fuera del alcance del análisis				

NULOS

tabla	columna	valor total	valor nulos
user_info	user_id	36000	NO HAY
user_info	age	36000	NO HAY
user_info	sex	36000	NO HAY

user_info	last_month_salary	36000	7199
user_info	number_dependents	36000	943
loans_outstanding	loan_id	305335	NO HAY
loans_outstanding	user_id	305335	NO HAY
loans_outstanding	loan_type	305335	NO HAY
loans_detail	user_id	36000	NO HAY
loans_detail	more_90_days_overdue	36000	NO HAY
loans_detail	using_lines_not_secured_personal_assets	36000	NO HAY
loans_detail	number_times_delay_payment_loan_30_59_days	36000	NO HAY
loans_detail	debt_ratio	36000	NO HAY
loans_detail	number_times_delay_payment_loan_60_89_days	36000	NO HAY
default	user_id	36000	NO HAY
default	default_flag	36000	NO HAY

DUPLICADOS

TABLA	COLUMNA	DUPLICADOS
user_info	user_id	NO HAY
loans_outstanding	loan_id	NO HAY
loans_detail	user_id	NO HAY
default	user_id	NO HAY

Selección de variables

En varios países está prohibido que los organismos financieros tomen decisiones basadas en variables discriminatorias como lo es el **género**, la etnia, la religión, etc. Por esta razón **no** se debe utilizar la variable género en el modelo.

Además de conocer las variables que **no** se pueden utilizar, también podemos usar la **correlación** para identificar variables que siguen un comportamiento similar, por ejemplo, si hay variables que aumentan proporcionalmente o inversamente.

La **alta correlación** entre dos variables indica que debemos elegir solo una de ellas para nuestro análisis, ya que si mantenemos las dos variables, puede ser que las reglas que


















vamos a crear más adelante le den un mayor peso a estas variables y esto se conoce como **multicolinealidad**.

La multicolinealidad generalmente ocurre cuando hay altas correlaciones entre dos o más variables predictoras . En otras palabras, una variable predictora se puede utilizar para predecir la otra. Esto crea información redundante, sesgando los resultados en un modelo de regresión.

Para ayudar a tomar la decisión de cuál de estas dos variables mantener para el modelo, podemos usar la **desviación estándar**.

Analizando los coeficientes de desviación estándar, podemos identificar cuál de las dos variables con alta correlación tiene una desviación mayor, lo que la hace más representativa y más interesante para el análisis, por lo que excluimos la variable con menor desviación.

COMPUERTAS AND & OR

FUNCIONES LÓGICAS BÁSICAS																																																																																							
NOMBRE	AND - Y	OR - O	XOR O-exclusiva	NOT Inversor	NAND	NOR																																																																																	
SÍMBOLO																																																																																							
SÍMBOLO																																																																																							
TABLA DE VERDAD	<table><tr><th>a</th><th>b</th><th>z</th></tr><tr><td>0</td><td>0</td><td>0</td></tr><tr><td>0</td><td>1</td><td>0</td></tr><tr><td>1</td><td>0</td><td>0</td></tr><tr><td>1</td><td>1</td><td>1</td></tr></table>	a	b	z	0	0	0	0	1	0	1	0	0	1	1	1	<table><tr><th>a</th><th>b</th><th>z</th></tr><tr><td>0</td><td>0</td><td>0</td></tr><tr><td>0</td><td>1</td><td>1</td></tr><tr><td>1</td><td>0</td><td>1</td></tr><tr><td>1</td><td>1</td><td>1</td></tr></table>	a	b	z	0	0	0	0	1	1	1	0	1	1	1	1	<table><tr><th>a</th><th>b</th><th>z</th></tr><tr><td>0</td><td>0</td><td>0</td></tr><tr><td>0</td><td>1</td><td>1</td></tr><tr><td>1</td><td>0</td><td>1</td></tr><tr><td>1</td><td>1</td><td>0</td></tr></table>	a	b	z	0	0	0	0	1	1	1	0	1	1	1	0	<table><tr><th>a</th><th>z</th></tr><tr><td>0</td><td>1</td></tr><tr><td>1</td><td>0</td></tr></table>	a	z	0	1	1	0	<table><tr><th>a</th><th>b</th><th>z</th></tr><tr><td>0</td><td>0</td><td>1</td></tr><tr><td>0</td><td>1</td><td>1</td></tr><tr><td>1</td><td>0</td><td>1</td></tr><tr><td>1</td><td>1</td><td>0</td></tr></table>	a	b	z	0	0	1	0	1	1	1	0	1	1	1	0	<table><tr><th>a</th><th>b</th><th>z</th></tr><tr><td>0</td><td>0</td><td>1</td></tr><tr><td>0</td><td>1</td><td>0</td></tr><tr><td>1</td><td>0</td><td>0</td></tr><tr><td>1</td><td>1</td><td>0</td></tr></table>	a	b	z	0	0	1	0	1	0	1	0	0	1	1	0
a	b	z																																																																																					
0	0	0																																																																																					
0	1	0																																																																																					
1	0	0																																																																																					
1	1	1																																																																																					
a	b	z																																																																																					
0	0	0																																																																																					
0	1	1																																																																																					
1	0	1																																																																																					
1	1	1																																																																																					
a	b	z																																																																																					
0	0	0																																																																																					
0	1	1																																																																																					
1	0	1																																																																																					
1	1	0																																																																																					
a	z																																																																																						
0	1																																																																																						
1	0																																																																																						
a	b	z																																																																																					
0	0	1																																																																																					
0	1	1																																																																																					
1	0	1																																																																																					
1	1	0																																																																																					
a	b	z																																																																																					
0	0	1																																																																																					
0	1	0																																																																																					
1	0	0																																																																																					
1	1	0																																																																																					
EQUIVALENTE EN CONTACTOS																																																																																							
AXIOMA	$z = a \cdot b$	$z = a + b$	$z = \bar{a} \cdot b + a \cdot \bar{b}$	$z = \bar{a}$	$z = \bar{a} \cdot \bar{b}$	$z = \overline{a + b}$																																																																																	

U (user_info)	D (default)	OR
NO	NO	✗
SI	NO	✓
NO	SI	✓
SI	SI	✓

Con el or al menos una de las dos tiene que ser **Verdad** por lo que únicamente no va a traer resultados cuando ambas sean **falsas**.

En el siguiente Query uní las tablas **default** y **user_id**. Posteriormente utilicé un **Where** que me trajera los clientes **NO morosos** y después utilice un **OR** para que me trajera aquellos valores donde el salario **NO** es nulo. Des esta manera aseguro que mi Query me elimine

únicamente los 130 registros donde los clientes son morosos y además no tengan valores de salario.

```
SELECT
U.*, D.default_flag
FROM `lucid-vector-408016.super_caja_ds.user_info` AS U
LEFT JOIN
`super_caja_ds.default` AS D
ON
U.user_id = D.user_id
WHERE default_flag!=1 #Donde default flag sea diferente de 1
OR last_month_salary IS NOT NULL #0 el salario NO sea nulo
```

Ya que anteriormente lo había utilizado de la siguiente manera y me quitaba casi 8mil valores.

```
WHERE NOT default_flag=1 AND last_month_salary IS NULL
```

Pero, el **WHERE** solo afecta a la primera condición, por lo que el **WHERE NOT** únicamente iba afectar al default_flag=1 y además que me trajera todo donde es nulo, porque el **AND** te va a regresar todo aquello donde ambas condiciones sean verdaderas.

U (user_info)	D (default)	AND
NO (no es nulo)	NO (es moroso)	✗
SI (es nulo)	NO (es moroso)	✗
NO (no es nulo)	SI (no es moroso)	✗
SI (es nulo)	SI (no es moroso)	✓

Son 683 clientes morosos (default_flag=1)
Faltan 7199 registros en último salario (last_month_salary = NULL)