# **Ficha técnica:** Proyecto 3 de Análisis de Datos

## **Título del proyecto:** Riesgo relativo

## Objetivo:

Armar un score crediticio a partir de un análisis de datos y la 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 el crédito, reduciendo así el riesgo de préstamos no reembolsables.

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

## Equipo:

Individual

## Herramientas y Tecnologías:

- Google BigQuery
- Google Colab
- Google Slides
- Google Looker Studio

## Lenguajes:

- SQL
- Python

## Insumos:

**dataset**

El conjunto de datos contiene datos sobre préstamos concedidos a un grupo de clientes del banco.

Divididos en 4 tablas:

- Tabla 1 user_info: con datos del usuario/cliente.
- Tabla 2 loans_outstanding (préstamos pendientes): con datos del tipo de préstamo.
- Tabla 3 loans_details: con el comportamiento de pago de estos préstamos.
- Tabla 4 default: con la identificación de clientes ya identificados como morosos.

## Diccionario de datos:

### Tabla 1:

| | | |
|---|---|---|
| user_ info | 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 |

### Tabla 2:

| | | |
|---|---|---|
| loans_ outstanding | 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) |

### Tabla 3:

| | | |
|---|---|---|
| loans_ detail | 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. Ratio de deuda = Deudas / 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) |

**Tabla 4:**

| | |
|---|---|
| default 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) |

## Procesamiento y análisis:

### HITO 1: Cálculo del riesgo relativo asociado a cada variable

### 2.1 Procesar y preparar base de datos

**1. Conectar/importar datos a otras herramientas**

- Se creó el proyecto3-riesgo-relativo-lab y el conjunto de datos Dataset en BigQuery.
- Tablas importadas: user_info, loans_outstanding, loans_details y default.

**2. Identificar y manejar valores nulos**

- Se identifican valores nulos a través de comandos SQL COUNT, WHERE y IS NULL.
- **user_info: 7199** valores nulos en la columna **last_month_salary** y **number_dependents.**
    - Unión de la base de datos user_info y default usando LEFT JOIN: se encontró que de **7199 valores nulos**, **7069 (19.64%)** clientes son buenos pagadores y **130 (0.36%)** son malos pagadores. Los datos nulos (7199) representan el 20% del total (36,000).
    - Con los comandos AVG, WHERE y GROUP BY, se calculó el promedio a la variable last_month_salary para cada categoría de cliente (buen pagador/mal pagador), sin considerar datos outliers, es decir, salarios mayores a 400,000.
    - Con los comandos IFNULL, CASE, WHEN, THEN, ELSE, se **IMPUTARON** los valores nulos de la variable last_month_salary colocando el promedio por categoría.
    - Con los comandos WITH, RANK se calculó la moda para la variable number_dependents para cada categoría de cliente (buen pagador/mal pagador).
    - Con los comandos IFNULL, CASE, WHEN, THEN, ELSE, se **IMPUTARON** los valores nulos de la variable number_dependents colocando la moda por categoría.
- **loans_outstanding:** 0 valores nulos.
- **loans_details:** 0 valores nulos.
- **default:** 0 valores nulos.

**3. Identificar y manejar valores duplicados**

- Se identifican duplicados a través de comandos SQL COUNT, GROUP BY, HAVING.
- **user_info:** no hay valores duplicados.
- **loans_outstanding:** no hay valores duplicados.
- **loans_details:** no hay valores duplicados.
- **default:** no hay valores duplicados.

**4. Identificar y manejar datos fuera del alcance del análisis**

- Se manejan variables que no son útiles para el análisis a través de comandos SQL SELECT EXCEPT.
- Se excluye la variable sex de la tabla user_info.
- Con el comando CORR y STDDEV, se calcula la correlación y la desviación estándar entre las variables more_90_days_overdue y number_times_delayed_payment_loan_30_59_days, y more_90_days_overdue y number_times_delayed_payment_loan_60_89_days. Se identifican las variables con alta correlación.
    - more_90_days_overdue y number_times_delayed_payment_loan_60_89_days tienen la correlación más alta con 0.9921

- Number_times_delayed_payment_loan_60_89_days tiene la desviación estándar más baja 4.1055.
- Para el análisis y limpieza se utiliza la variable more_90_days_overdue.

**5. Identificar y manejar datos inconsistentes en variables categóricas**

- Con los comandos DISTINCT, COUNT, CASE, WHEN, THEN, ELSE, LOWER se estandarizaron los datos de la variable loan_type.

**6. Identificar y manejar datos inconsistentes en variables numéricas**

- Con los comandos WITH, APPROX_QUANTILES, CASE, WHEN, ELSE, WHERE, se identifican los datos outliers de las tablas user_info y de loans_detail. Se utilizó la metodología de rango intercuartil.

| user_info | | | |
|---|---|---|---|
| **Variable** | **age** | **las_month_salary** | **number_dependents** |
| **Outliers** | 10 | 1187 | 3230 |

| loan_detail | |
|---|---|
| **Variable** | **Outliers** |
| **more_90_days_overdue** | 1946 |
| **using_lines_not_secured_personal_assets** | 177 |
| **number_times_delayed_payment_loan_30_59_days** | 5812 |
| **debt_ratio** | 7570 |
| **number_times_delayed_payment_loan_60_89_days** | 1865 |

**7. Crear nuevas variables**

- Con los comandos DISTINCT, SUM, CASE, WHEN, GROUP BY, se hizo una tabla agrupada por usuario, con una fila para cada cliente, mostrando el tipo de préstamo y la cantidad total.

**8. Unir tablas**

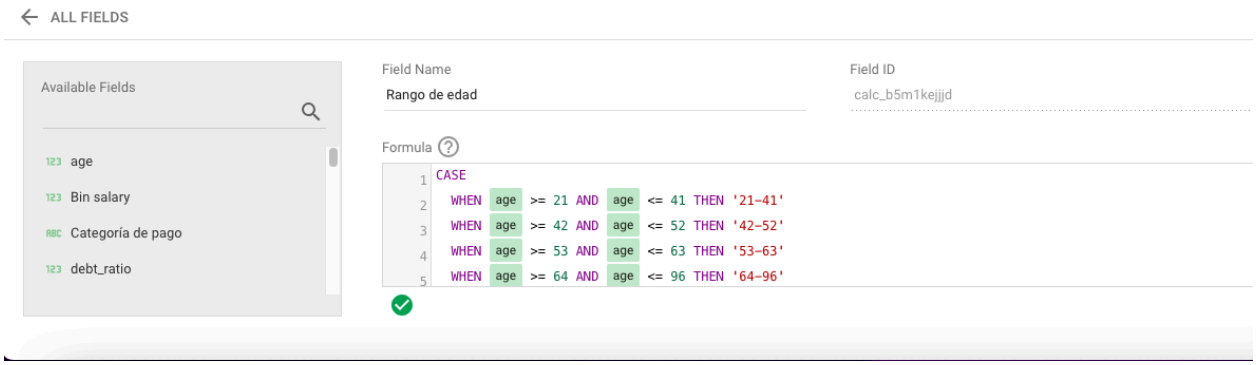
- Con el comando INNER JOIN se unieron las vistas user_default_limpia, loans_out_totales, loans_detail_limpia, creando una tabla consolidada con las variables limpias join_user_out_detail_limpia.

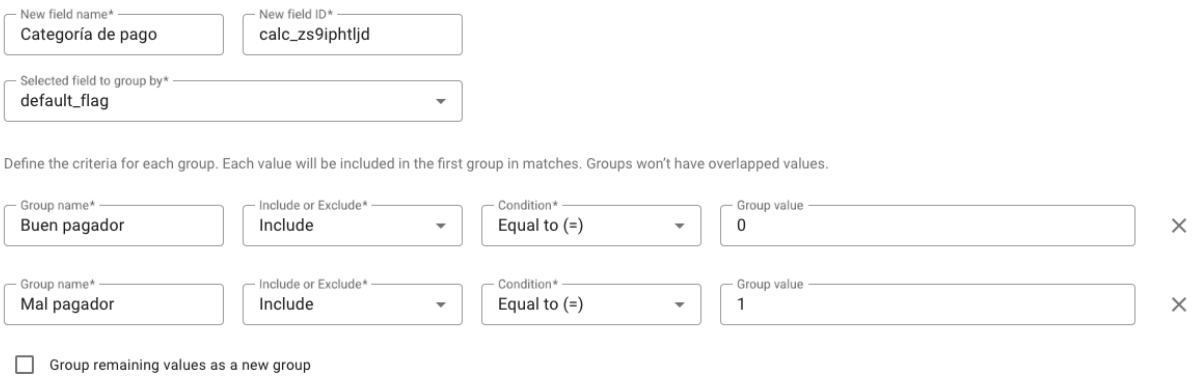
**2.2 Análisis exploratorio**

**9. Agrupar datos según variables categóricas**

- Se conectaron los datos a looker studio desde BigQuery.
- Se creó un campo calculado en looker studio para crear una clasificación de datos por rango de edad, utilizando los comandos CASE, WHEN, THEN y de está manera hacer un análisis exploratorio.

- Se creó un grupo categoría de pago para buen pagador y mal pagador de acuerdo a las flags 0 y 1.




## 10. Visualizar las variables categóricas

- Se realizaron gráficos de barras para la visualización de variables y exploración de datos en looker studio.

## 11. Aplicar medidas de tendencia central y aplicar medidas de dispersión

- Se crearon tablas en looker studio con las medidas de tendencia central (mediana, promedio) para comparar los datos por rango de edad y por categoría de pago.
- Se crearon tablas en looker studio con las medidas de dispersión (rango, desviación estándar) para comparar los datos por rango de edad y por categoría de pago.

## 12. Visualizar distribución

- Se crearon box plot para visualizar la distribución de las variables por rango de edad y categoría de pago en looker studio.
- Se realizaron box plot e histogramas para las variables en google colab usando python.

## 13. Aplicar correlación entre las variables numéricas

- Se creó una matriz de correlación de todas las variables en google colab utilizando Python.
- Con el comando CORR se calculó la correlación entre variables en BigQuery.

| Row | correlation_age_sala | correlation_age_mor | correlation_age_total | correlation_salay_tot |
|---|---|---|---|---|
| 1 | 0.068444825483... | -0.07137834102... | 0.135537786978... | 0.186380286614... |

**14. Calcular cuartiles, deciles o percentiles**

- Con los comandos WITH, NTILE, COUNT, GROUP BY, MIN, MAX, JOIN, se calcularon los cuartiles de cada variable, se contabilizó el número de usuarios por cuartil, el total de malos pagadores y se calculó el rango de cada cuartil.

    - **Age**

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTION

| Row | cuartiles_age | total_usuarios | total_malos_pagador | min_age | max_age |
|---|---|---|---|---|---|
| 1 | 1 | 8893 | 268 | 21 | 42 |
| 2 | 2 | 8892 | 194 | 42 | 52 |
| 3 | 3 | 8892 | 112 | 52 | 63 |
| 4 | 4 | 8892 | 47 | 63 | 109 |

    - **last_month_salary**

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTION GR

| Row | cuartiles_last_month | total_count | total_malos_pagador | min_last_month_sala | max_last_month_sal |
|---|---|---|---|---|---|
| 1 | 1 | 8893 | 245 | 0.0 | 3943.0 |
| 2 | 2 | 8892 | 275 | 3944.0 | 6600.0 |
| 3 | 3 | 8892 | 29 | 6600.0 | 7491.0 |
| 4 | 4 | 8892 | 72 | 7491.0 | 150000.0 |

    - **number_dependents**

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTIC

| Row | cuartiles_dependent | total_usuarios | total_malos_pagador | min_number_depend | max_number_depen |
|---|---|---|---|---|---|
| 1 | 1 | 8893 | 193 | 0 | 0 |
| 2 | 2 | 8892 | 91 | 0 | 0 |
| 3 | 3 | 8892 | 140 | 0 | 1 |
| 4 | 4 | 8892 | 197 | 1 | 13 |

- **debt_ratio**

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTI

| Row | cuartiles_debt_ratio | total_count | total_malos_pagador | min_debt_ratio | max_debt_ratio |
|---|---|---|---|---|---|
| 1 | 1 | 8893 | 123 | 0.0 | 0.181227143 |
| 2 | 2 | 8892 | 134 | 0.181248368 | 0.369299802 |
| 3 | 3 | 8892 | 203 | 0.369317294 | 0.881365417 |
| 4 | 4 | 8892 | 161 | 0.8814448 | 307001.0 |

- **using_lines_not_secured_personal_assets**

| Row | cuartiles_using_lines | total_usuarios | total_malos_pagador | min_using_lines | max_using_lines |
|---|---|---|---|---|---|
| 1 | 1 | 8893 | 8 | 0.0 | 0.028823045 |
| 2 | 2 | 8892 | 1 | 0.028829832 | 0.144330007 |
| 3 | 3 | 8892 | 29 | 0.144361012 | 0.529282357 |
| 4 | 4 | 8892 | 583 | 0.52937284 | 22000.0 |

- **number_times_delayed_payment_loan_30_59_days**

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTI

| Row | cuartiles_30_59 | total_count | total_malos_pagador | min_30_59_days | max_30_59_days |
|---|---|---|---|---|---|
| 1 | 1 | 8893 | 25 | 0 | 0 |
| 2 | 2 | 8892 | 4 | 0 | 0 |
| 3 | 3 | 8892 | 26 | 0 | 0 |
| 4 | 4 | 8892 | 566 | 0 | 11 |

- **more_90_days_overdue**

| Row | cuartiles_90 | total_count | total_malos_pagador | min_90 | max_90 |
|---|---|---|---|---|---|
| 1 | 1 | 8893 | 17 | 0 | 0 |
| 2 | 2 | 8892 | 7 | 0 | 0 |
| 3 | 3 | 8892 | 15 | 0 | 0 |
| 4 | 4 | 8892 | 582 | 0 | 15 |

- **total_loans**

| Row | cuartiles_total_loans | total_count | total_malos_pagador | min_total_loans | max_total_loans |
|---|---|---|---|---|---|
| 1 | 1 | 8893 | 257 | 1 | 5 |
| 2 | 2 | 8892 | 157 | 5 | 8 |
| 3 | 3 | 8892 | 100 | 8 | 11 |
| 4 | 4 | 8892 | 107 | 11 | 57 |

**2.3 Aplicar técnica de análisis**

**15. Calcular riesgo relativo**

- Con los comando WITH, NTILE, COUNT, MIN, MAX, CASE, WHEN, LEFT JOIN, se calculó el riesgo relativo en BigQuery para las variables, obteniendo una tabla con los cuartiles, total de usuarios, total de malos y buenos pagadores, riesgo relativo, y el rango de los cuartiles.

**Hito 2: Crear un puntaje (score) para los clientes a través del análisis del hito 1.**

**3.1 Procesar y preparar la base de datos**

**17. Crear nuevas variables**

- Con el comando IF se crearon variables dummies para las 7 variables seleccionadas en el análisis.

**3.2 Aplicar técnica de análisis**

**18. Segmentación de clientes**

- Con los comandos WITH, CASE, WHEN, ELSE se calculó un score por user_id, sumando los valores de las 7 variables dummies seleccionadas para el análisis y obteniendo un puntaje.
- Con el puntaje obtenido se clasificó a los usuarios como *buen pagador* si su puntaje era <=3 y, >4 como *mal pagador*, el puntaje máximo es 7.
- En google colab se creó una matriz de confusión para evaluar qué tan buena es la clasificación.

**HITO 3: Regresión logística**

**4.1 Procesar y preparar base de datos**

**19. Conectar/importar datos a otras herramientas**

- Se creó un nuevo notebook en google colab.
- Se importaron las librerías, se realizó autenticación y conexión a BigQuery.
- Se hizo la autenticación y creación de cliente BigQuery.
- Se definió el proyecto, dataset y tabla.

- Se creó la consulta con datos que necesitamos para el dataframe.
- Se creó el dataframe.
- Se imprimieron las primeras filas del dataframe.

**20. Crear nuevas variables**

- Se convirtieron las etiquetas de clasificación a valores numéricos utilizando comandos de python.

**4.2 Aplicar técnica de análisis**

**21. Regresión logística**

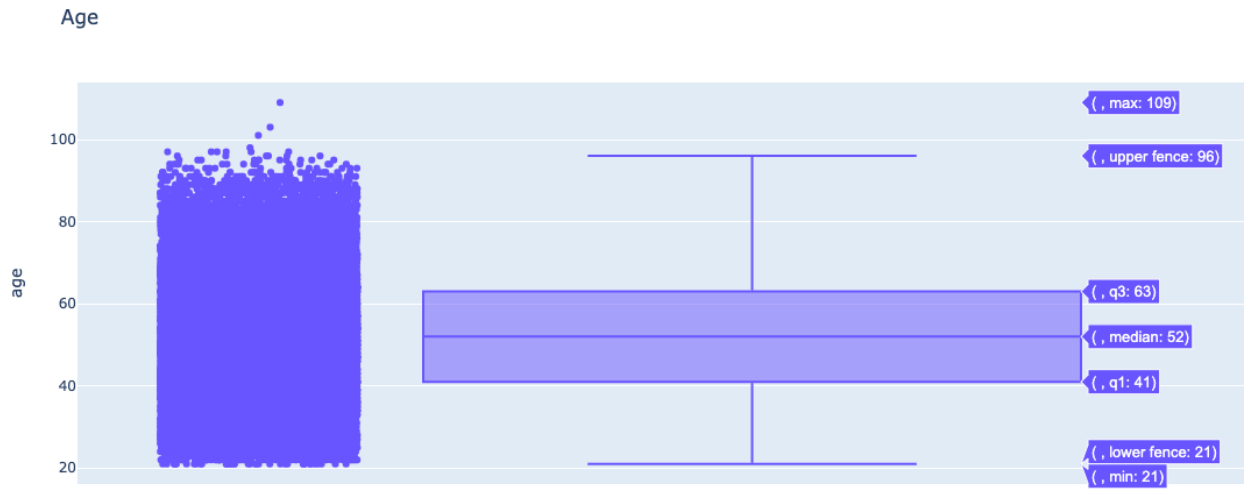
- Se importaron las librerías dentro de google colab.
- Se convirtieron las variables categóricas en variables numéricas.
- Se asignaron las variables al eje *x* y *y*, se creó el modelo y entrenó el modelo.
- Se imprimieron los parámetros de *w* y *b*, se creó un rango de valores para la variable independiente, se calcularon las probabilidades.
- Se imprimió el modelo de regresión, la curva y los puntos.

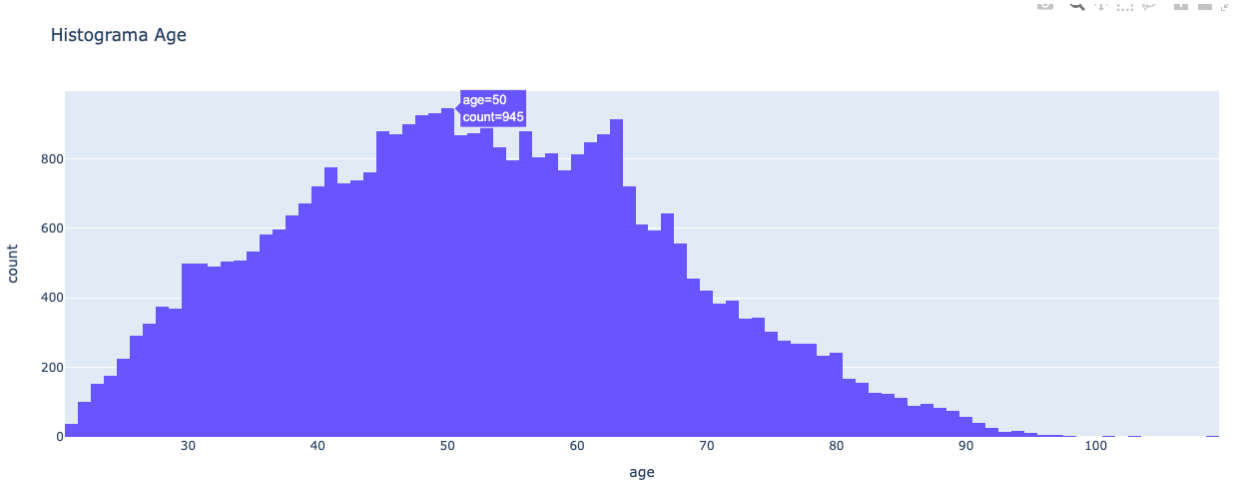
## Resultados y Conclusiones:

**1. Manejo de datos outliers**

- Se realizaron box plots e histogramas interactivos en google colab usando python para visualizar mejor los resultados encontrados, adicional se hicieron nuevas consultas para encontrar los valores más extremos un top 30 y top 70, concluyendo lo siguiente:

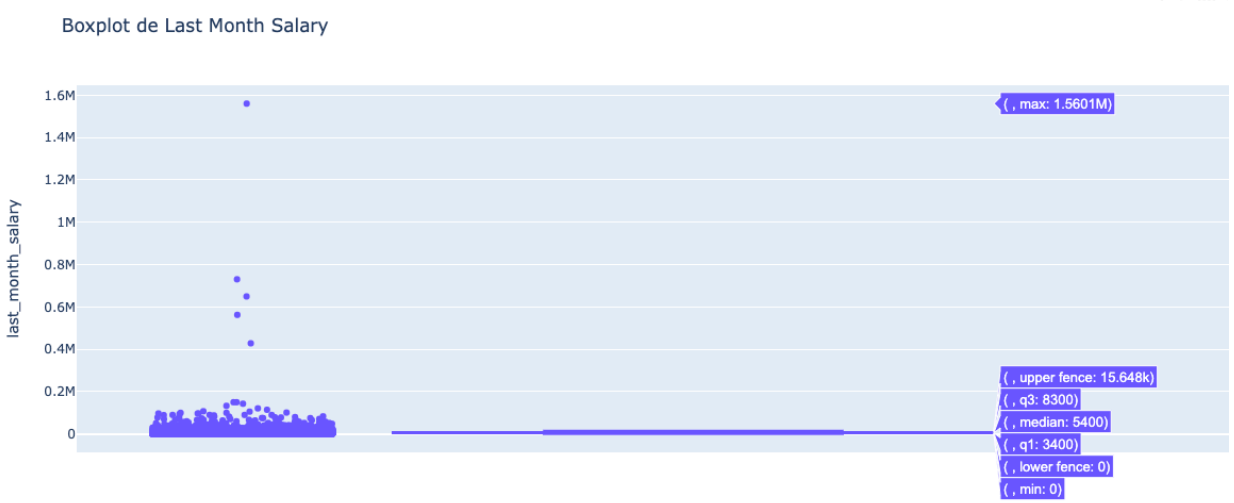
    - **Age**: se eliminan de la base 10 datos outliers.



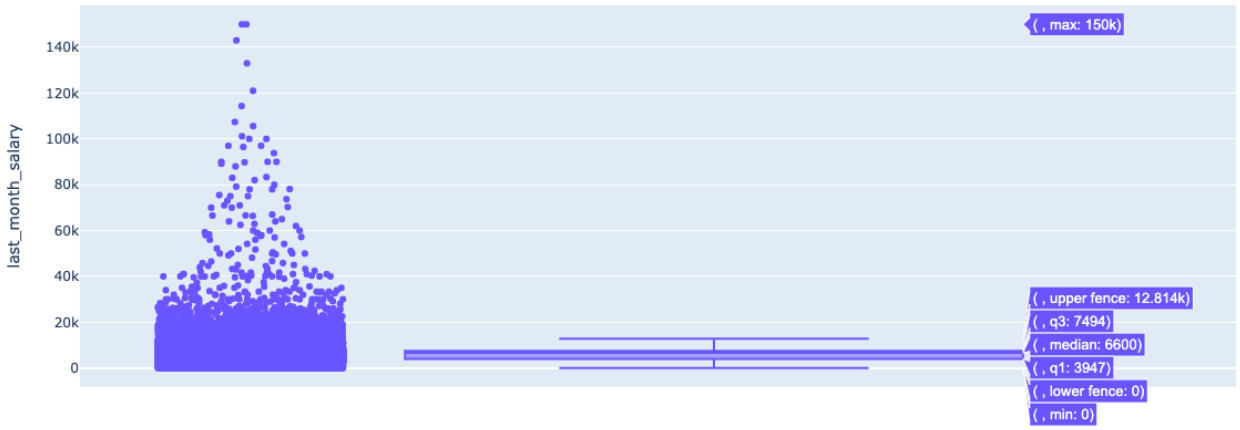



| Row | user_id | age |
|---|---|---|
| 1 | 135 | 109 |
| 2 | 26810 | 103 |
| 3 | 6586 | 101 |
| 4 | 1276 | 98 |
| 5 | 33063 | 97 |
| 6 | 19667 | 97 |
| 7 | 13661 | 97 |
| 8 | 17248 | 97 |
| 9 | 14729 | 97 |

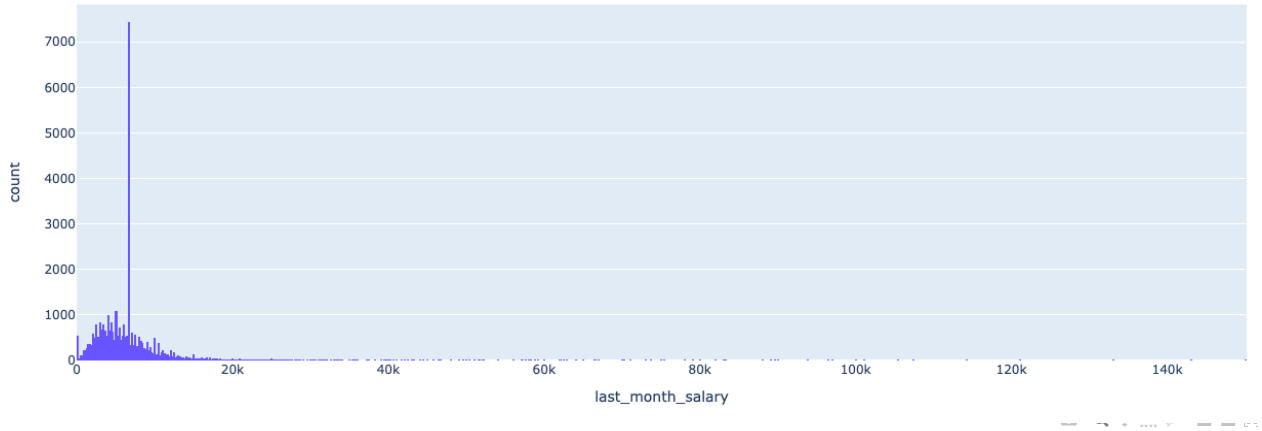
- **Last_month_salary**: en los gráficos y en los datos se observan 5 valores muy por encima de los demás, por lo que se descartaran los registros con salarios por arriba de 400,000.



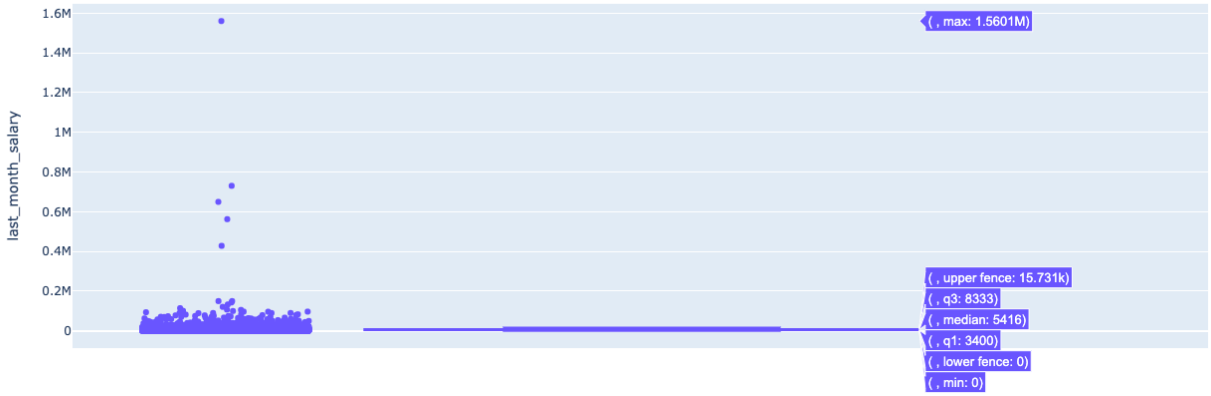
Flag 1 last_month_salary

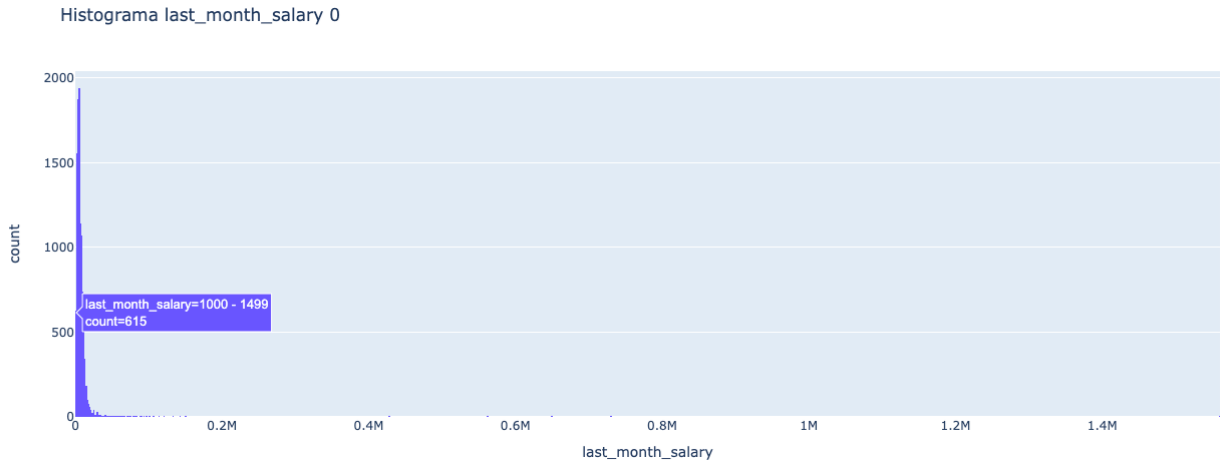



Histograma FLag1 last_month_salary



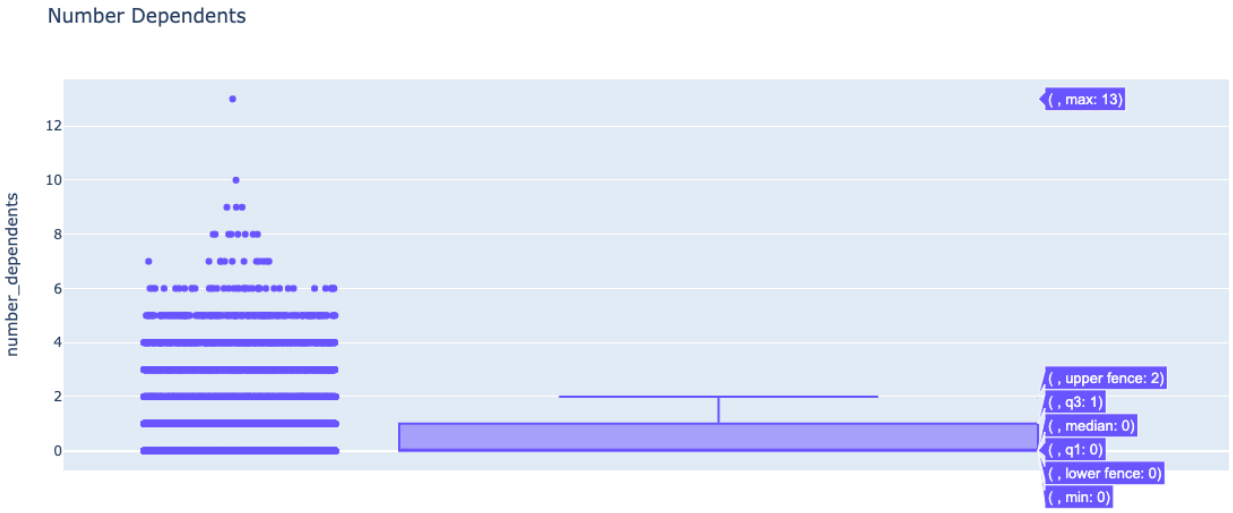

last_month_salary 0

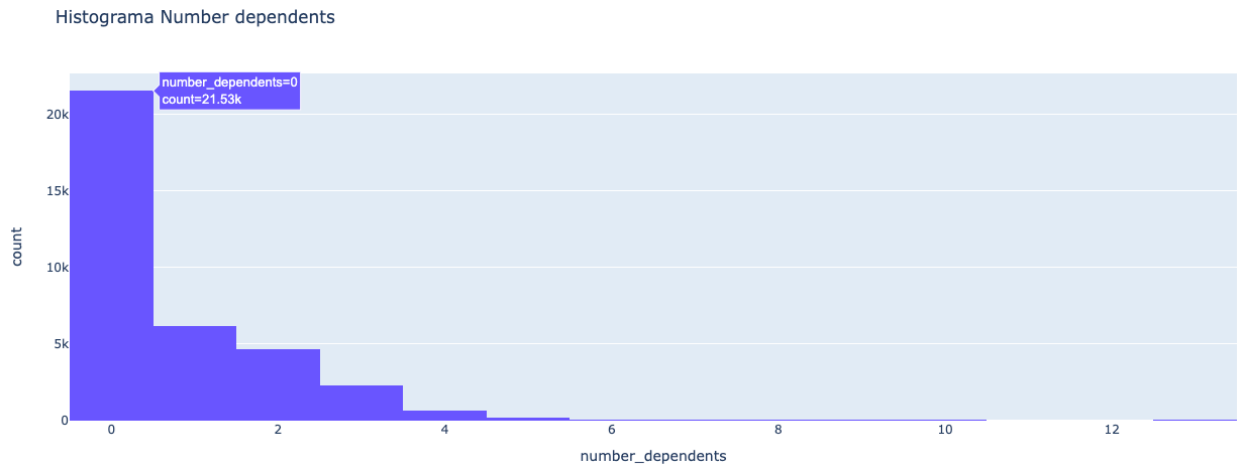





| Row | user_id | age | sex | last_month_salary |
|---|---|---|---|---|
| 1 | 21096 | 44 | F | 1560100.0 |
| 2 | 6543 | 67 | M | 730483.0 |
| 3 | 23384 | 49 | F | 649587.0 |
| 4 | 22076 | 61 | F | 562466.0 |
| 5 | 24043 | 43 | M | 428250.0 |
| 6 | 15932 | 48 | M | 150000.0 |

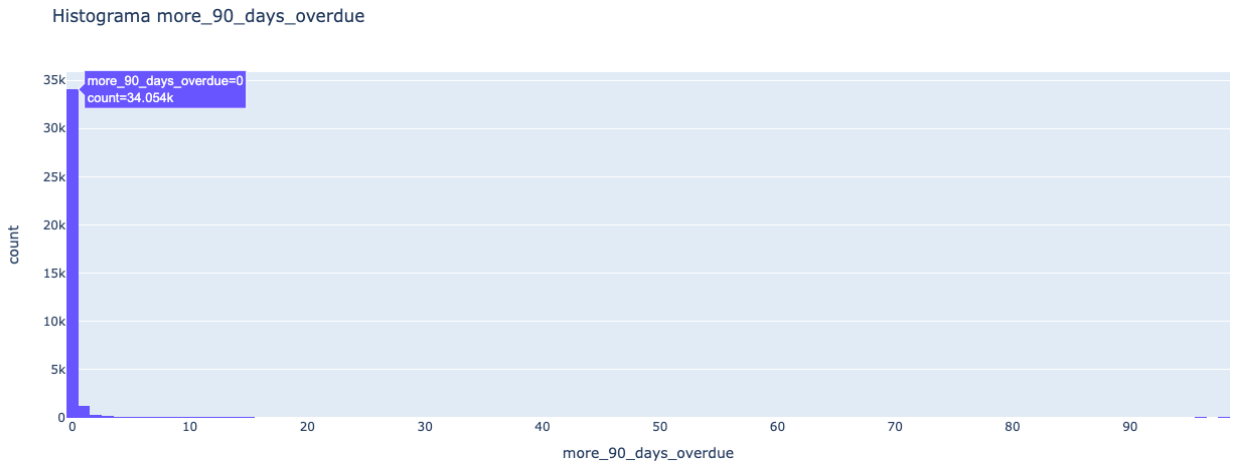
- **number_dependents**: en las gráficas y en los datos se observa un valor muy alejado, se decide mantener este dato.






| Row | ser_id | age | sex | last_month_salary | number_dependents |
|---|---|---|---|---|---|
| 1 | 15517 | 53 | M | 3333.0 | 13 |
| 2 | 14692 | 47 | M | 9166.0 | 10 |
| 3 | 34884 | 48 | F | 11400.0 | 9 |
| 4 | 22582 | 48 | M | 16666.0 | 9 |
| 5 | 12123 | 37 | M | 3300.0 | 9 |

- **More_90_days_overdue:** en los gráficos y en los datos se observan 63 valores muy por encima de los demás (98 y 96) y con datos inconsistentes en las otras variables, por lo que se descartaran los registros mayores a 20.



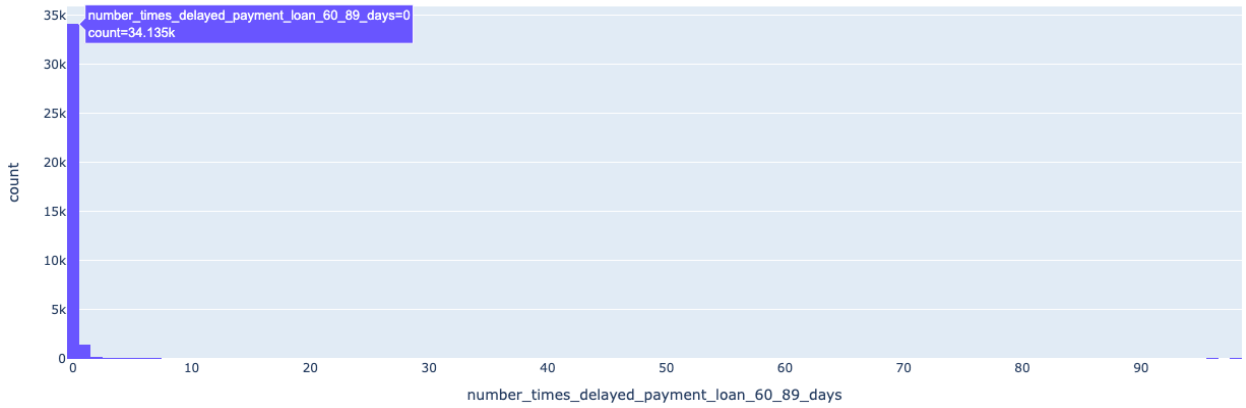

| Row | user_id | more_90_days_overd | using_lines_not_secu | number_times_delay | debt_ratio | number_times_delay |
|---|---|---|---|---|---|---|
| 1 | 4256 | 98 | 0.9999999 | 98 | 0.0 | 98 |
| 2 | 16129 | 98 | 0.9999999 | 98 | 0.0 | 98 |
| 3 | 29590 | 98 | 0.9999999 | 98 | 0.0 | 98 |
| 4 | 8840 | 98 | 0.9999999 | 98 | 9.0 | 98 |
| 5 | 27941 | 98 | 0.9999999 | 98 | 0.037367158 | 98 |
| 6 | 17431 | 98 | 0.9999999 | 98 | 0.0 | 98 |
| 7 | 2341 | 98 | 0.9999999 | 98 | 0.0 | 98 |
| 8 | 31126 | 98 | 0.9999999 | 98 | 0.012993503 | 98 |
| 9 | 5323 | 98 | 0.9999999 | 98 | 0.0 | 98 |
| 10 | 21979 | 98 | 0.9999999 | 98 | 0.0 | 98 |

- **Number_times_deleyed_payment_loan_30_59_days:** en los gráficos y en los datos se observan 63 valores muy por encima de los demás (98 y 96) y con datos inconsistentes en las otras variables, por lo que se descartaran los registros mayores a 20.


| Row | user_id | more_90_days_overd | using_lines_not_secu | number_times_delay | debt_ratio | number_times_delay |
|---|---|---|---|---|---|---|
| 1 | 27941 | 98 | 0.9999999 | 98 | 0.037367158 | 98 |
| 2 | 20647 | 98 | 0.9999999 | 98 | 0.0 | 98 |
| 3 | 29725 | 98 | 0.9999999 | 98 | 0.0 | 98 |
| 4 | 33138 | 98 | 0.9999999 | 98 | 0.0 | 98 |
| 5 | 12771 | 98 | 0.9999999 | 98 | 0.0 | 98 |
| 6 | 22269 | 98 | 0.9999999 | 98 | 89.0 | 98 |
| 7 | 22882 | 98 | 0.9999999 | 98 | 0.0 | 98 |
| 8 | 32628 | 98 | 0.9999999 | 98 | 0.0 | 98 |
| 9 | 30818 | 98 | 0.9999999 | 98 | 0.0 | 98 |
| 10 | 12542 | 98 | 0.9999999 | 98 | 0.0 | 98 |

- **Number_times_deleyed_payment_loan_60_89_days:** en los gráficos y en los datos se observan 63 valores muy por encima de los demás (98 y 96) y con datos inconsistentes en las otras variables, por lo que se descartaran los registros mayores a 20.

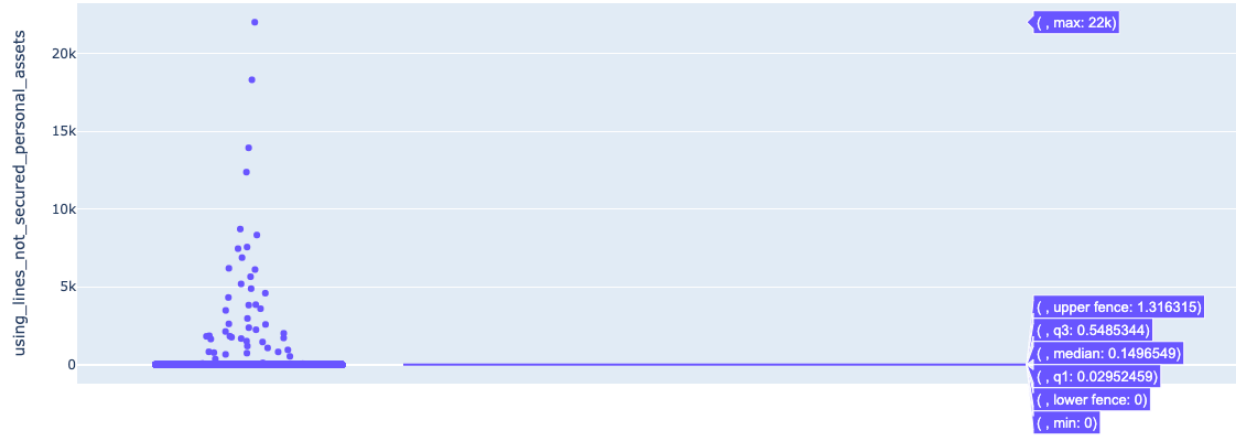
Histograma number_times_delayed_payment_loan_60_89_days

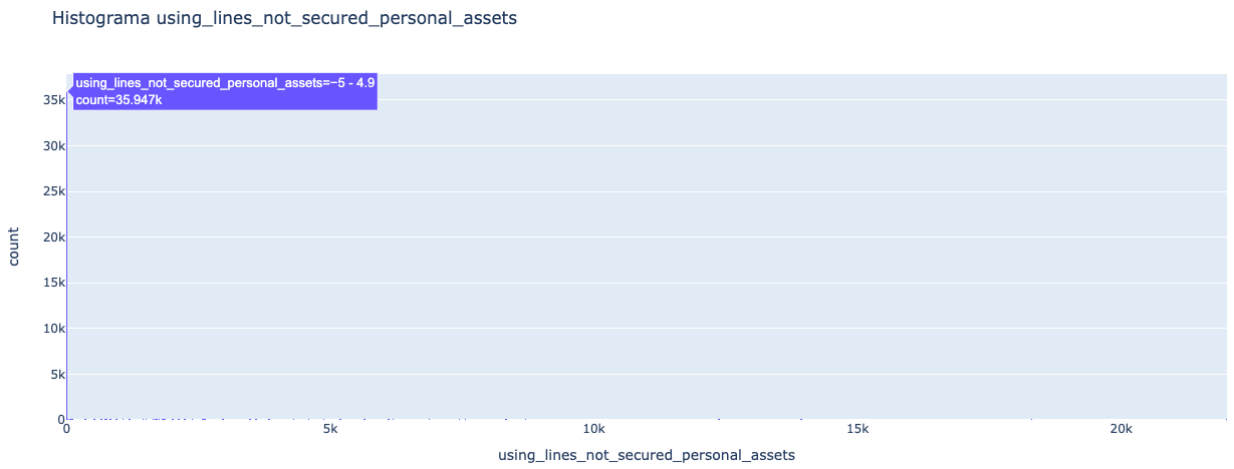



| Row | user_id | more_90_days_overd | using_lines_not_secu | number_times_delay | debt_ratio | number_times_delay |
|---|---|---|---|---|---|---|
| 3 | 12542 | 98 | 0.9999999 | 98 | 0.0 | 98 |
| 4 | 6787 | 98 | 0.9999999 | 98 | 0.00757815 | 98 |
| 5 | 20647 | 98 | 0.9999999 | 98 | 0.0 | 98 |
| 6 | 25244 | 98 | 0.9999999 | 98 | 54.0 | 98 |
| 7 | 23057 | 98 | 0.9999999 | 98 | 0.0 | 98 |
| 8 | 30272 | 98 | 0.9999999 | 98 | 0.0 | 98 |
| 9 | 4256 | 98 | 0.9999999 | 98 | 0.0 | 98 |
| 10 | 12021 | 98 | 0.9999999 | 98 | 0.0 | 98 |

- **Using_lines_not_secured**: se observan 4 valores por encima de los demás.

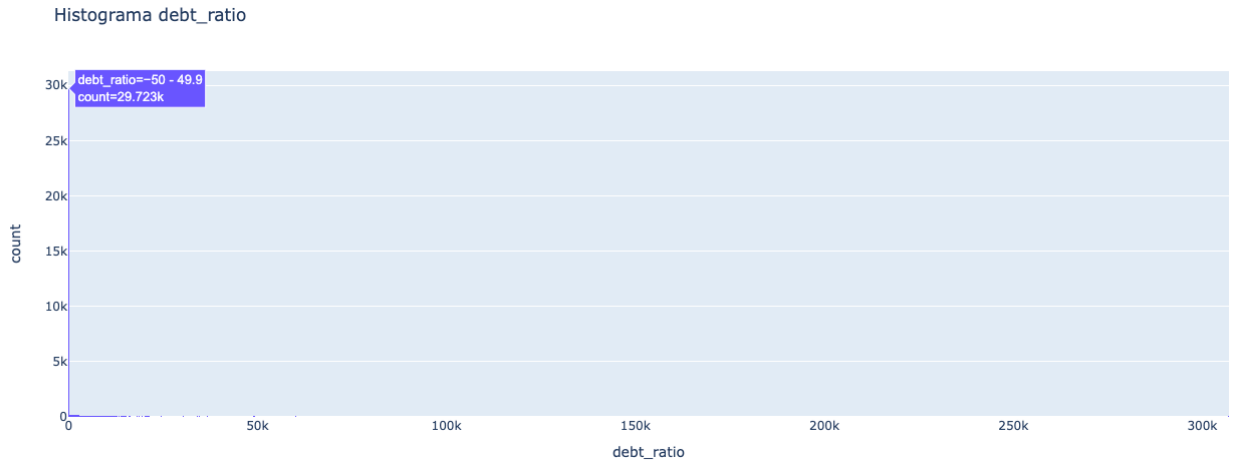
using_lines_not_secured_personal_assets






| Row | user_id | more_90_days_overd | using_lines_not_secu | number_times_delay | debt_ratio |
|---|---|---|---|---|---|
| 1 | 10876 | 0 | 22000.0 | 0 | 1.080020131 |
| 2 | 6579 | 0 | 18300.0 | 0 | 0.221582273 |
| 3 | 3680 | 0 | 13930.0 | 0 | 4902.0 |
| 4 | 15531 | 0 | 12369.0 | 2 | 0.134352002 |
| 5 | 1410 | 0 | 8710.0 | 0 | 0.442809151 |

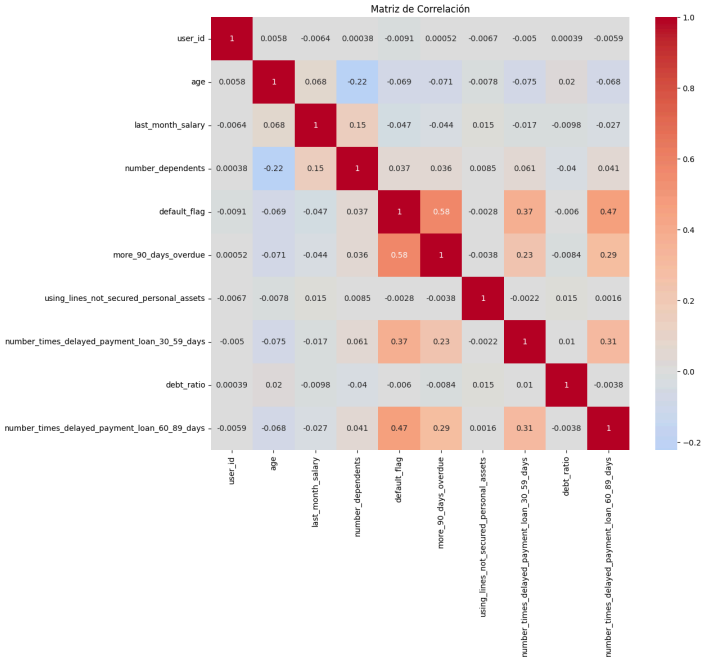
- **debt_ratio**: se observa un valor por encima de los demás por lo que se elimina de la base de datos.




| Row | user_id | more_90_days_overd | using_lines_not_secu | number_times_delay | debt_ratio | number_times_delay | debt_ratio_status |
|---|---|---|---|---|---|---|---|
| 1 | 18739 | 0 | 0.034289594 | 1 | 307001.0 | 0 | Above Upper Bound |
| 2 | 13321 | 0 | 0.351736756 | 0 | 60212.0 | 1 | Above Upper Bound |
| 3 | 8057 | 0 | 0.895728527 | 2 | 49112.0 | 0 | Above Upper Bound |
| 4 | 8821 | 0 | 0.0 | 0 | 36705.0 | 0 | Above Upper Bound |
| 5 | 24713 | 0 | 0.104444633 | 0 | 34719.0 | 0 | Above Upper Bound |
| 6 | 32026 | 0 | 0.843508169 | 2 | 34102.0 | 1 | Above Upper Bound |
| 7 | 26959 | 0 | 0.002067762 | 0 | 30295.0 | 0 | Above Upper Bound |
| 8 | 2424 | 0 | 0.012106977 | 0 | 24591.0 | 0 | Above Upper Bound |
| 9 | 34835 | 0 | 8328.0 | 0 | 21395.0 | 1 | Above Upper Bound |
| 10 | 18571 | 0 | 0.012259994 | 2 | 20948.0 | 0 | Above Upper Bound |
| 11 | 22690 | 0 | 0.897636745 | 2 | 20351.0 | 0 | Above Upper Bound |
| 12 | 11398 | 0 | 0.408641659 | 0 | 20243.0 | 0 | Above Upper Bound |
| 13 | 11926 | 0 | 0.223530906 | 0 | 19378.0 | 0 | Above Upper Bound |

## 2. Matriz de correlación de variables

- Se creó una matriz para identificar las variables que están más correlacionadas.



### 3. Cálculo del riesgo relativo para las variables

- **age**
  - Cuartil 1 y 2: valores con el riesgo relativo mayor a 1, hay un mayor riesgo de incumplimiento de pago en estos rangos de edad, siendo el grupo de 21 a 42 años el más probable de ser un mal pagador.
  - Cuartil 3 y 4: valores menor a 1, hay menor riesgo de incumplimiento en el grupo expuesto (mal pagador) en comparación con el no expuesto (buen pagador).

| Row | age_quartile | total_count | total_bad_payers | total_good_payers | riesgo_relativo | min_age | max_age |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 8890 | 268 | 8622 | 2.277449596420... | 21 | 42 |
| 2 | 2 | 8890 | 193 | 8697 | 1.352702290718... | 42 | 52 |
| 3 | 3 | 8889 | 113 | 8776 | 0.667372883232... | 52 | 63 |
| 4 | 4 | 8889 | 47 | 8842 | 0.245663022417... | 63 | 96 |

- **last_month_salary**
  - Cuartil 1 y 2: valores mayores a 1 en el riesgo relativo, hay un mayor riesgo de incumplimiento de pago en estos rangos de salarios, los usuarios con el rango salarial de 3948 a 6600 son los más probables de ser malos pagadores.
  - Cuartil 3 y 4: valores menor a 1, hay menor riesgo de incumplimiento en el grupo expuesto (mal pagador) en comparación con el no expuesto (buen pagador).

| Row | salary_quartile | total_count | total_bad_payers | total_good_payers | riesgo_relativo | min_salary | max_salary |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 8890 | 245 | 8645 | 1.954640643323... | 0.0 | 3947.0 |
| 2 | 2 | 8890 | 275 | 8615 | 2.384214256454... | 3948.0 | 6600.0 |
| 3 | 3 | 8889 | 29 | 8860 | 0.146970481281... | 6600.0 | 7494.0 |
| 4 | 4 | 8889 | 72 | 8817 | 0.393472130778... | 7495.0 | 150000.0 |

- **number_dependents**
  - Cuartil 1 y 4: valores mayores a 1 en el riesgo relativo, hay un mayor riesgo de incumplimiento en el grupo expuesto en comparación con el no expuesto. Es más probable que los usuarios con hijos sean malos pagadores.
  - Cuartil 2 y 3: valores menor a 1, hay menor riesgo de incumplimiento en el grupo expuesto (mal pagador) en comparación con el no expuesto (buen pagador).

| Row | dependents_quartile | total_count | total_bad_payers | total_good_payers | riesgo_relativo | min_dependents | max_dependents |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 8890 | 193 | 8697 | 1.352702290718... | 0 | 0 |
| 2 | 2 | 8890 | 91 | 8799 | 0.515055712375... | 0 | 0 |
| 3 | 3 | 8889 | 140 | 8749 | 0.873246360927... | 0 | 1 |
| 4 | 4 | 8889 | 197 | 8692 | 1.393972463315... | 1 | 13 |

- **more_90_days_overdue**
    - Cuartil 1, 2 y 3: valores menor a 1, hay menor riesgo de incumplimiento en el grupo expuesto (mal pagador) en comparación con el no expuesto (buen pagador).
    - Cuartil 4: valor mayor a 1, hay un mayor riesgo de incumplimiento en el grupo expuesto en comparación con el no expuesto. Es más probable que los usuarios que se retrasen una sola vez más de 90 días sean malos pagadores.

| Row | overdue_quartile | total_count | total_bad_payers | total_good_payers | riesgo_relativo | min_overdue | max_overdue |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 8890 | 17 | 8873 | 0.084430754102... | 0 | 0 |
| 2 | 2 | 8890 | 7 | 8883 | 0.034199389571... | 0 | 0 |
| 3 | 3 | 8889 | 15 | 8874 | 0.074262994979... | 0 | 0 |
| 4 | 4 | 8889 | 582 | 8307 | 44.77258841956... | 0 | 15 |

- **total_loans**
    - Cuartil 1: valor de riesgo relativo mayor a 1, hay mayor riesgo de incumplimiento de pago en los usuarios que tienen de 1 a 5 préstamos activos.
    - Cuartil 2: no hay una diferencia significativa entre el grupo expuesto y el no expuesto.
    - Cuartil 3 y 4: valores menor a 1, hay menor riesgo de incumplimiento en el grupo expuesto (mal pagador) en comparación con el no expuesto (buen pagador).

| Row | loans_quartile | total_count | total_bad_payers | total_good_payers | riesgo_relativo | min_loans | max_loans |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 8890 | 257 | 8633 | 2.117973028096... | 1 | 5 |
| 2 | 2 | 8890 | 157 | 8733 | 1.015010084946... | 5 | 8 |
| 3 | 3 | 8889 | 100 | 8789 | 0.575858924604... | 8 | 11 |
| 4 | 4 | 8889 | 107 | 8782 | 0.624560456612... | 11 | 57 |

- **using_lines_not_secured_personal_assets**
    - Cuartil 1, 2 y 3: valores menor a 1, hay menor riesgo de incumplimiento en el grupo expuesto (mal pagador) en comparación con el no expuesto (buen pagador).
    - Cuartil 4: valor de riesgo relativo mayor a 1, hay mayor riesgo de incumplimiento de pago en los usuarios que tienen un uso de su línea de crédito personal de bienes no asegurados mayor a 0.53.

| Row | unsecured_quartile | total_count | total_bad_payers | total_good_payers | riesgo_relativo | min_unsecured | max_unsecured |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 8890 | 8 | 8882 | 0.039148776875... | 0.0 | 0.028848558 |
| 2 | 2 | 8890 | 1 | 8889 | 0.004838346819... | 0.028851958 | 0.144448889 |
| 3 | 3 | 8889 | 29 | 8860 | 0.146970481281... | 0.144482624 | 0.529476034 |
| 4 | 4 | 8889 | 583 | 8306 | 46.02976772000... | 0.529492733 | 22000.0 |

- **debt_ratio**

- Cuartil 1 y 2: valores menor a 1, hay menor riesgo de incumplimiento en el grupo expuesto (mal pagador) en comparación con el no expuesto (buen pagador).
- Cuartil 3: valor de riesgo relativo mayor a 1, hay mayor riesgo de incumplimiento de pago en los usuarios que tienen un ratio de deuda mayor a 0.37.

| Row | debt_quartile | total_count | total_bad_payers | total_good_payers | riesgo_relativo | min_debt_ratio | max_debt_ratio |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 8890 | 123 | 8767 | 0.740908290078... | 0.0 | 0.181303116 |
| 2 | 2 | 8890 | 134 | 8756 | 0.825400110407... | 0.18130803 | 0.369363482 |
| 3 | 3 | 8889 | 203 | 8686 | 1.457047068012... | 0.369388628 | 0.881365417 |
| 4 | 4 | 8889 | 161 | 8728 | 1.050078749015... | 0.8814448 | 60212.0 |

**4. Validación de hipótesis**

**Hipótesis:**

- **Los más jóvenes tienen un mayor riesgo de impago.**

**Conclusión:** Se válida la hipótesis, el rango de edad con mayor riesgo relativo (2.28) es de 21 a 42 años, cuartil 1.

- **Las personas con más cantidad de préstamos activos tienen mayor riesgo de ser malos pagadores.**

**Conclusión:** Se refuta está hipótesis, el rango de préstamos activos con mayor riesgo relativo (2.12) es de 1 a 5 préstamos corresponden al cuartil 1.

- **Las personas que han retrasado sus pagos por más de 90 días tienen mayor riesgo de ser malos pagadores.**

**Conclusión:** Se válida la hipótesis, el valor máximo de veces que un cliente se retrasa es de 15 con un riesgo relativo (44.78) correspondiente al cuartil 4.

**5. Perfil de cliente con mayor probabilidad de ser mal pagador**

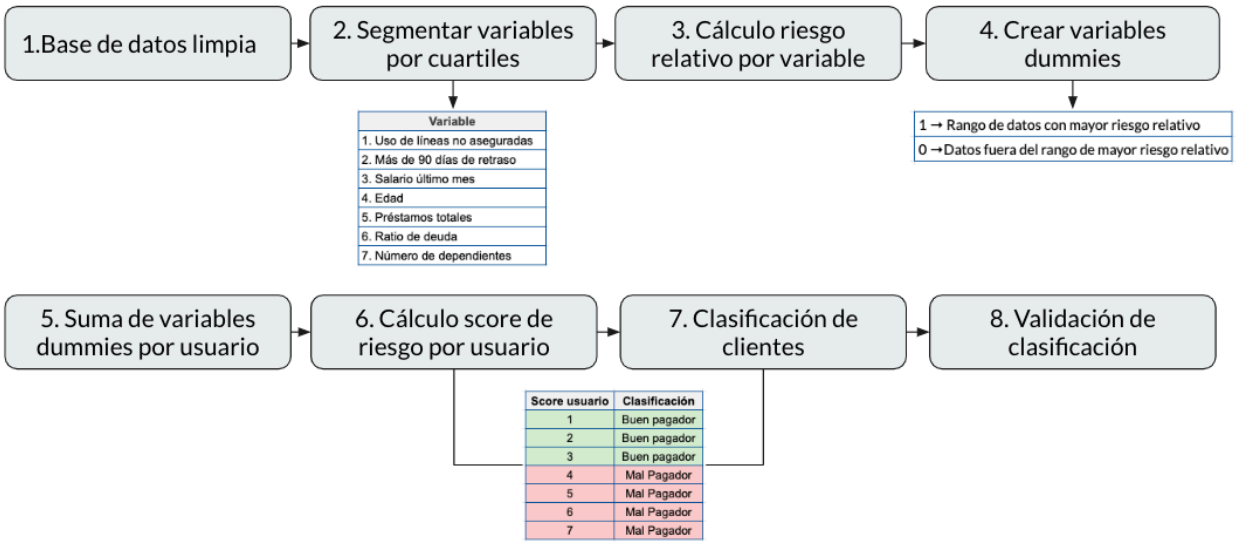
SUPER CAJA

**Tabla con el rango de datos de cada variable que tiene mayor riesgo de ser mal pagadora**

| Variable | Cuartiles | Mínimo | Máximo | Total malos pagad... | Total buenos paga... | Total usuarios | Riesgo Relativo |
|---|---|---|---|---|---|---|---|
| age | 1 | 21 | 42 | 268 | 8,622 | 8,890 | 2.28 |
| debt_ratio | 3 | 0.37 | 0.88 | 203 | 8,686 | 8,889 | 1.46 |
| last_month_salary | 2 | 3,948 | 6,600 | 275 | 8,615 | 8,890 | 2.38 |
| more_90_days | 4 | 0 | 15 | 582 | 8,307 | 8,889 | 44.77 |
| number_dependents | 4 | 1 | 13 | 197 | 8,692 | 8,889 | 1.39 |
| ratio_credito | 4 | 0.53 | 22,000 | 583 | 8,306 | 8,889 | 46.03 |
| total_loans | 1 | 1 | 5 | 257 | 8,633 | 8,890 | 2.12 |

**Conclusión:** El usuario con mayor probabilidad de ser un mal pagador tiene entre 21 - 42 años, su salario está entre 3948 - 6600 dolares, tiene al menos 1 dependiente, tiene de 1 a 5 préstamos activos, posee un ratio de crédito por arriba de 0.53 y un debt ratio mayor a 0.37.

### 6. Score crediticio




**Resultados**

- **Buenos pagadores:** 28,571.
- **Malos pagadores:** 6,987.

| Score Total ▲ | Usuarios | % |
|---|---|---|
| 0 | 3,691 | 10.38% |
| 1 | 7,968 | 22.41% |
| 2 | 9,314 | 26.19% |
| 3 | 7,598 | 21.37% |
| 4 | 4,662 | 13.11% |
| 5 | 1,807 | 5.08% |
| 6 | 466 | 1.31% |
| 7 | 52 | 0.15% |

### 7. Matriz de confusión

**Valores:**

- **True Positives (TP)**: 28,486 (Buen Pagador predicho correctamente como Buen Pagador)
- **False Positives (FP)**: 6,451 (Buen Pagador predicho incorrectamente como Mal Pagador)
- **False Negatives (FN)**: 85 (Mal Pagador predicho incorrectamente como Buen Pagador)
- **True Negatives (TN)**: 536 (Mal Pagador predicho correctamente como Mal Pagador)

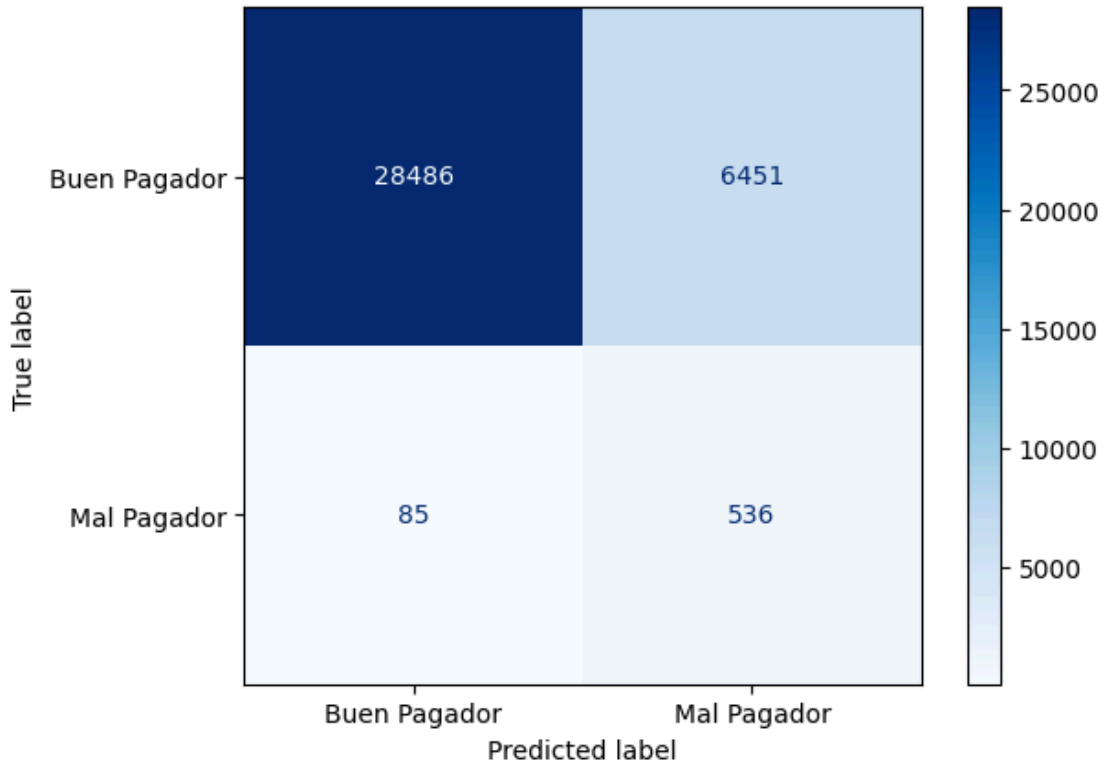
**Reporte de Clasificación**

- **Buen Pagador**:
    - **Precisión:** 1.00 (100%) – De todos los casos predichos como Buen Pagador, el 100% son realmente Buen Pagador.
    - **Recall:** 0.82 (82%) – De todos los casos que son realmente Buen Pagador, el 82% fueron correctamente identificados.
    - **F1-score**: 0.90 (90%) – La media armónica de la precisión y el recall.
    - **Support**: 34,937 – La cantidad de muestras reales de Buen Pagador.
- **Mal Pagador**:
    - **Precisión**: 0.08 (8%) – De todos los casos predichos como Mal Pagador, solo el 8% son realmente Mal Pagador.
    - **Recall**: 0.86 (86%) – De todos los casos que son realmente Mal Pagador, el 86% fueron correctamente identificados.
    - **F1-score**: 0.14 (14%) – La media armónica de la precisión y el recall.
    - **Support**: 621 – La cantidad de muestras reales de Mal Pagador.

**Métricas Globales**

- **Accuracy (Precisión Global)**: 0.82 (82%) – El porcentaje de todas las predicciones que son correctas.
- **Macro Average**:
    - **Precisión**: 0.54 (54%) – La media no ponderada de la precisión para cada clase.
    - **Recall**: 0.84 (84%) – La media no ponderada del recall para cada clase.
    - **F1-score**: 0.52 (52%) – La media no ponderada del F1-score para cada clase.
- **Weighted Average**:
    - **Precisión**: 0.98 (98%) – La media ponderada de la precisión para cada clase.
    - **Recall**: 0.82 (82%) – La media ponderada del recall para cada clase.
    - **F1-score**: 0.88 (88%) – La media ponderada del F1-score para cada clase.

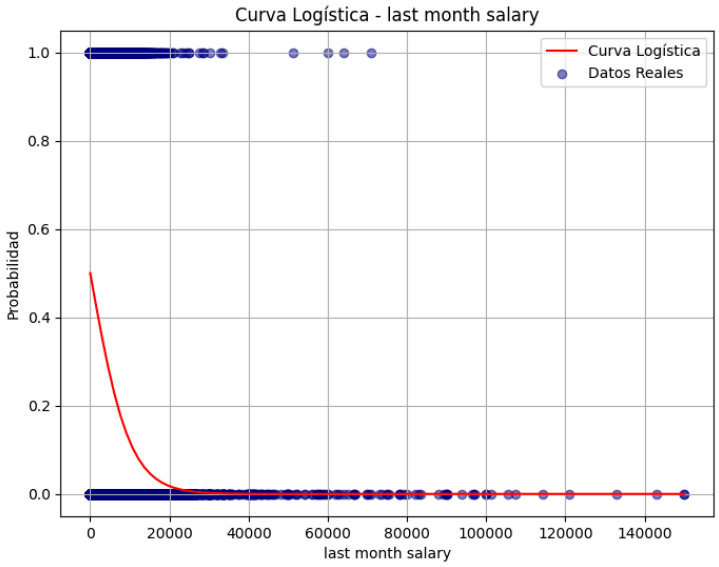
```
Matriz de Confusión:
[[28486  6451]
 [   85   536]]
Reporte de Clasificación:
              precision    recall  f1-score   support

Buen Pagador       1.00      0.82      0.90     34937
 Mal Pagador       0.08      0.86      0.14       621

    accuracy                           0.82     35558
   macro avg       0.54      0.84      0.52     35558
weighted avg       0.98      0.82      0.88     35558
```




**8. Regresión logística**

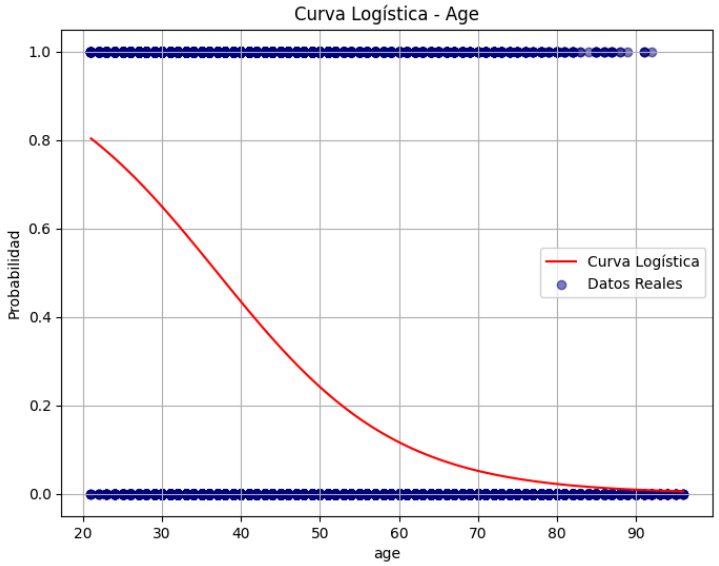
**Last month salary**

- **Salarios bajos:** la probabilidad de que el evento ocurra es alta, es decir, salarios por debajo de los 5,000 dólares tienen mayor probabilidad de incumplimiento de pago.
- **Salarios altos:** a medida que el salario aumenta la probabilidad de incumplimiento de pago disminuye.



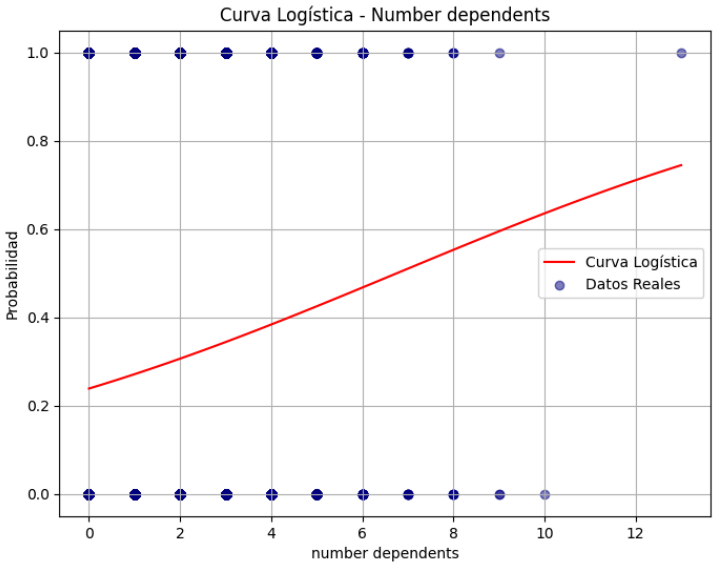

**Age**

- La probabilidad de incumplimiento es más alta para los usuarios más jóvenes y disminuye a medida que la edad aumenta.



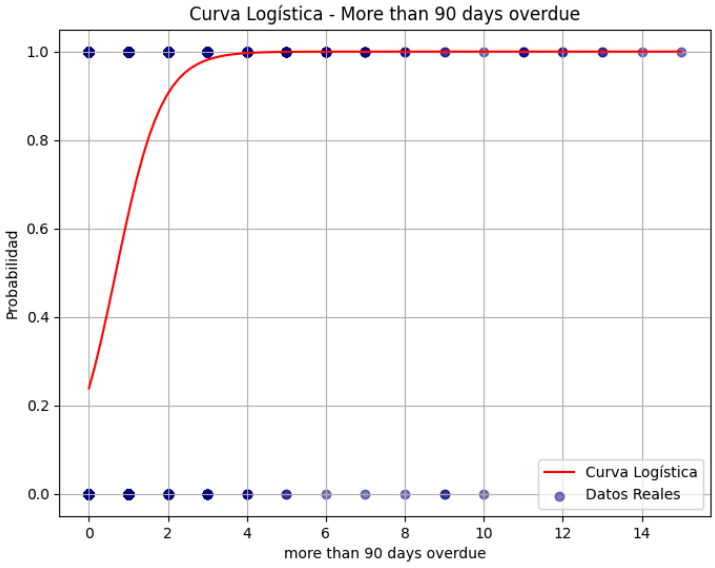

**Number dependents**

- De acuerdo al modelo, tener más dependientes aumenta la probabilidad de incumplimiento de pago, lo que puede deberse a que un mayor número de dependientes podría significar más responsabilidades financieras y, por lo tanto, un mayor riesgo de incumplimiento.



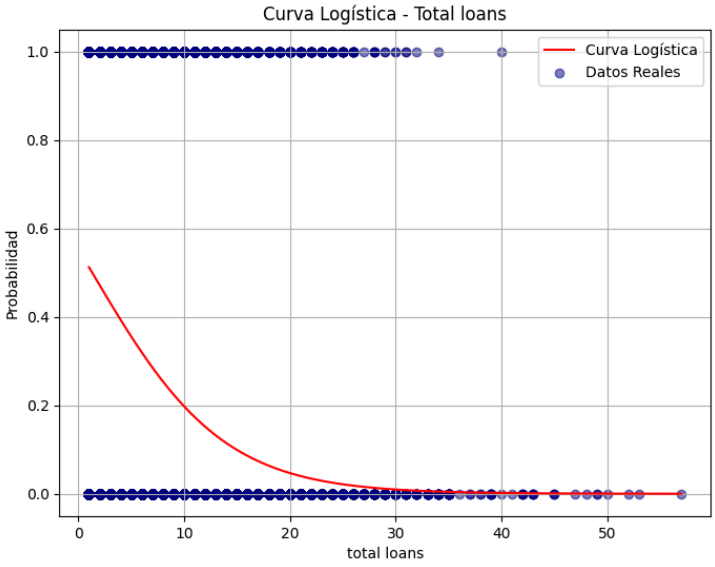

**More than 90 days overdue**

- El modelo sugiere que si una persona se ha retraso por más de 90 días una vez o más, la probabilidad de incumplimiento es extremadamente alta, casi segura. Esto se refleja en la forma de la curva, que se estabiliza en una probabilidad cercana a 1 después de pocos incidentes de retraso.
- **El número de veces que una persona ha estado en mora por más de 90 días es un predictor muy fuerte de incumplimiento de pago.** El modelo muestra que con solo uno o dos incidentes de mora de más de 90 días, la probabilidad de incumplimiento casi se asegura. Los datos reales refuerzan esta conclusión, con un claro patrón de incumplimiento conforme aumenta el número de incidentes de mora.



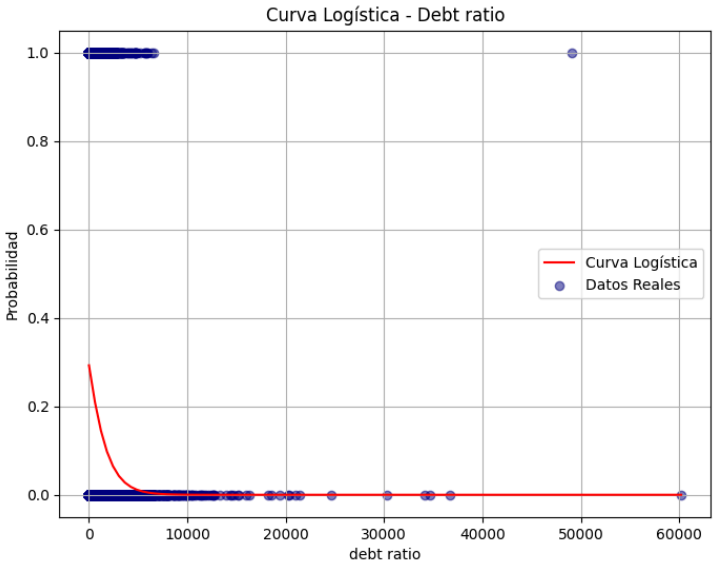
**Total loans**

- La curva de regresión logística muestra una tendencia decreciente. A medida que aumenta el número total de préstamos, la probabilidad de incumplimiento de pago disminuye. Esto sugiere que los clientes con más préstamos tienen una menor probabilidad de incumplir sus pagos.



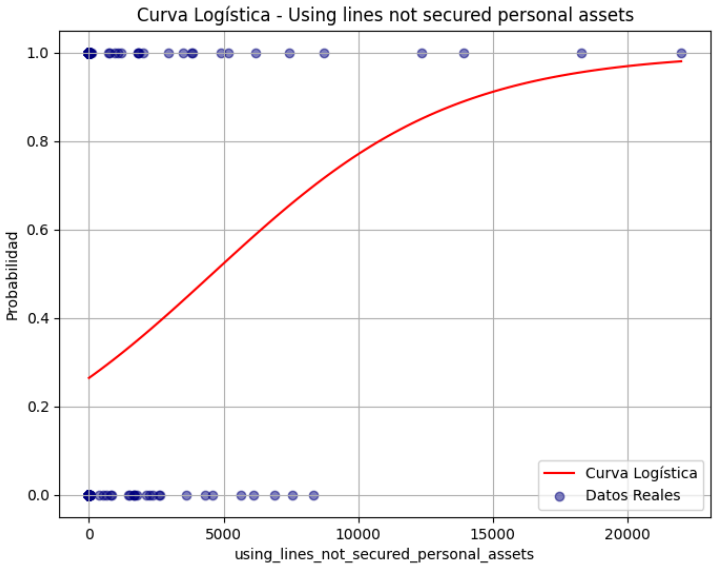

**Debt ratio**

- La curva indica que, a medida que aumenta el debt ratio, la probabilidad de incumplimiento disminuye considerablemente.
- Debt ratios bajos (cerca de 0), la probabilidad de incumplimiento es muy alta.
- Este patrón podría implicar que los clientes que tienen un mayor debt ratio están administrando bien sus deudas o tienen una mayor capacidad financiera para manejar sus obligaciones.



**Using lines not secured personal assets**

- A valores bajos de uso de líneas no aseguradas (cerca de 0), la probabilidad de incumplimiento es muy baja, casi cercana a 0.
- A medida que el uso de estas líneas de crédito supera los 5,000 unidades monetarias, la probabilidad de incumplimiento comienza a aumentar de manera más significativa.
- Para montos que se acercan o superan los 20,000 unidades monetarias, la probabilidad de incumplimiento se aproxima al 100%. Esto sugiere que los clientes que utilizan grandes cantidades de crédito no asegurado son extremadamente propensos a incumplir.




## Recomendaciones:

- **Política de Evaluación de Edad:** Implementar criterios más estrictos para la aprobación de préstamos a clientes jóvenes, especialmente en el rango de 21 a 42 años. Esto podría incluir mayores requisitos o tasas de interés más altas.

- **Ajustes en la Política de Préstamos Basados en Ingresos:** Establecer límites más conservadores para los préstamos a clientes con ingresos bajos, o considerar un enfoque de préstamos escalonados que ofrezca montos menores y tasas más altas a estos segmentos.

- **Consideración del Número de Dependientes:** Incorporar el número de dependientes en el proceso de evaluación de riesgos. Los clientes con más dependientes podrían requerir un análisis financiero más detallado antes de la aprobación del préstamo.

- **Historial de Mora:** Utilizar el historial de mora como un criterio clave en la decisión de otorgar un préstamo. Los clientes con cualquier historial de mora de más de 90 días deberían ser considerados de alto riesgo y, posiblemente, rechazados para nuevos préstamos.

- 
- **Control de Número de Préstamos:** Implementar un límite en la cantidad de préstamos activos que un cliente puede tener. Exigir una revisión exhaustiva de la capacidad de pago antes de aprobar nuevos préstamos a clientes con varios préstamos existentes.
- 
- **Revisión de Líneas de Crédito No Aseguradas:** Implementar políticas que limiten el uso de líneas de crédito no aseguradas, especialmente para clientes que ya tienen un alto nivel de endeudamiento no asegurado. Considerar la opción de exigir garantías o activos asegurados como condición para la concesión de nuevos créditos a estos clientes.
- 
- **Revisión de ratio de deuda:** Establecer umbrales más bajos para el ratio de deuda aceptable y denegar o limitar los préstamos a aquellos que excedan estos umbrales. Considerar ofrecer productos con montos menores y plazos más cortos a clientes con ratios de deuda elevados.

## Limitaciones/Próximos pasos:

- Mejorar la consideración de valores outliers en el análisis, es decir, si los gráficos e información me dicen que son datos atípicos, no considerarlos en el análisis.
- Ajustar el modelo de score, para tener mejorar las métricas de evaluación.

## Enlaces de intéres:

- **Github:**

https://github.com/Jessica-Cazares/Lab_proyecto3_riesgo_relativo/tree/main

- **Dashboard Looker studio:**

https://lookerstudio.google.com/reporting/57f2b28d-a385-4732-9f3c-10a62d8e279e