Credit Score: Un problema de clasificación

Mariano Giongrande
Comisión 61175

Índice

Introducción

El presente trabajo tiene como objetivo definir los lineamientos para el desarrollo de un caso de un problema de clasificación. A lo largo del presente se encontrarán definiciones sobre el alcance del mismo, las hipótesis planteadas, los datos con los que se va a intentar dar respuesta a esas hipótesis, los problemas encontrados en el dataset y sus posibles soluciones. Adicionalmente se harán comentarios respecto del análisis exploratorio, incluyendo gráficos que se encuentran en la respectiva notebook. Por último, conclusiones y recomendaciones para poder avanzar en la segunda y última etapa del proyecto.

Objetivo

El objetivo principal del presente trabajo es realizar un modelo de clasificación de riesgo crediticio, en tres categorías (Good, Standard y Poor) a partir de un dataset con datos de clientes que poseen líneas de crédito abiertas.

Hipótesis

Los principales drivers para realizar una exitosa calificación crediticia son los siguientes:

- Ingreso Anual (annual income)
- Saldo Mensual (monthly balance)
- Cuota Mensual Equivalente (Equated Monthly Installment - EMI)
- Comportamiento de pago (Payment behavior)
- Deuda pendiente (outstanding debt)
- Tasa de interés (interest rate)
- Cantidad de pagos atrasados (number of delayed payments)

Dataset

El dataset con el que se realiza el trabajo fue descargado de Kaggle, tiene 12.500 registros y 28 columnas (para detalle de los tipos de datos, ver tabla 1):

- *ID*: corresponde al ID del registro.
- *Customer_ID*: corresponde al ID del cliente.
- *Month*: corresponde al mes del registro.
- *Name*: corresponde al nombre del cliente.
- *Age*: corresponde a la edad del cliente.
- *SSN*: corresponde al número de seguridad social del cliente.

- *Occupation*: corresponde a la ocupación del cliente.
- *Annual_Income*: corresponde al ingreso anual del cliente.
- *Monthly_Inhand_Salary*: corresponde al ingreso neto mensual del cliente.
- *Num_Bank_Accounts*: corresponde a la cantidad de cuentas bancarias del cliente.
- *Num_Credit_Card*: corresponde a la cantidad de tarjetas de crédito que tiene un cliente.
- *Interest_Rate*: corresponde a la tasa de interés de la tarjeta de crédito.
- *Num_of_Loan*: corresponde a la cantidad de créditos tomados del banco.
- *Type_of_Loan*: corresponde al tipo de crédito tomado por el cliente.
- *Delay_from_due_date*: corresponde al promedio de días de demora respecto de la fecha de pago.
- *Num_of_Delayed_Payment*: corresponde al promedio de pagos demorados por el cliente.
- *Changed_Credit_Limit*: corresponde al cambio porcentual en el límite de la tarjeta de crédito.
- *Num_Credit_Inquiries*: corresponde a la cantidad de veces que fue consultado su historial crediticio en el período en cuestión.
- *Credit_Mix*: corresponde a la clasificación del mix de créditos.
- *Outstanding_Debt*: corresponde al saldo de deuda a pagar.
- *Credit_Utilization_Ratio*: corresponde al ratio de uso de la tarjeta de crédito.
- *Crédit_History_Age*: corresponde a la antigüedad del historial crediticio del cliente.
- *Payment_of_Min_Amount*: corresponde a la existencia o no de pago mínimo.
- *Total_EMI_per_month*: corresponde a la Cuota Equivalente Mensual pagada por el cliente (corresponde a la cuota fija pagada).
- *Amount_invested_monthly*: corresponde al monto invertido mensualmente por el cliente.
- *Payment_Behavior*: corresponde al comportamiento de pago del cliente.
- *Monthly_Balance*: corresponde al saldo mensual del cliente.
- *Credit_Score*: corresponde a la calificación crediticia.

```
RangeIndex: 12500 entries, 0 to 12499
Data columns (total 28 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   ID                        12500 non-null  object
 1   Customer_ID               12500 non-null  object
 2   Month                     12500 non-null  object
 3   Name                      11271 non-null  object
 4   Age                       12500 non-null  object
 5   SSN                       12500 non-null  object
 6   Occupation                12500 non-null  object
 7   Annual_Income             12500 non-null  object
 8   Monthly_Inhand_Salary     10584 non-null  float64
 9   Num_Bank_Accounts         12500 non-null  int64
 10  Num_Credit_Card           12500 non-null  int64
 11  Interest_Rate             12500 non-null  int64
 12  Num_of_Loan               12500 non-null  object
 13  Type_of_Loan              11074 non-null  object
 14  Delay_from_due_date       12500 non-null  int64
 15  Num_of_Delayed_Payment    11660 non-null  object
 16  Changed_Credit_Limit      12500 non-null  object
 17  Num_Credit_Inquiries      12243 non-null  float64
 18  Credit_Mix                12500 non-null  object
 19  Outstanding_Debt          12500 non-null  object
 20  Credit_Utilization_Ratio  12500 non-null  float64
 21  Credit_History_Age        11380 non-null  object
 22  Payment_of_Min_Amount     12500 non-null  object
 23  Total_EMI_per_month       12500 non-null  float64
 24  Amount_invested_monthly   11914 non-null  object
 25  Payment_Behaviour         12500 non-null  object
 26  Monthly_Balance           12353 non-null  object
 27  Credit_Score              12500 non-null  object
dtypes: float64(4), int64(4), object(20)
memory usage: 2.7+ MB
```

Tabla 1

Análisis Exploratorio

Este proceso consistió en analizar las distintas columnas para encontrar datos faltantes, datos mal ingresados (como ser palabras en lugares de números, caracteres en lugares de números, etc), convertir columnas a su tipo de dato correcto, y eliminar columnas innecesarias.

Para mayor detalle de lo realizado en cada columna se recomienda consultar la notebook correspondiente.

Una vez finalizado el proceso se llegó al dataset final según lo expuesto en la tabla 2.

```
RangeIndex: 6518 entries, 0 to 6517
Data columns (total 19 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       6518 non-null   int64
 1   Occupation                6518 non-null   object
 2   Annual_Income             6518 non-null   float64
 3   Num_Bank_Accounts         6518 non-null   int64
 4   Interest_Rate             6518 non-null   int64
 5   Num_of_Loan               6518 non-null   int64
 6   Delay_from_due_date       6518 non-null   int64
 7   Changed_Credit_Limit      6518 non-null   float64
 8   Outstanding_Debt          6518 non-null   float64
 9   Credit_Utilization_Ratio  6518 non-null   float64
 10  Payment_of_Min_Amount     6518 non-null   object
 11  Payment_Behaviour         6518 non-null   object
 12  Credit_Score              6518 non-null   object
 13  Num_Credit_Inquiries      6518 non-null   float64
 14  Credit_History            6518 non-null   float64
 15  Amount_invested_monthly   6518 non-null   float64
 16  Num_of_Delayed_Payment    6518 non-null   float64
 17  Month_Balance             6518 non-null   float64
 18  Total_EMI_per_month       6518 non-null   float64
dtypes: float64(10), int64(5), object(4)
memory usage: 967.6+ KB
```

Tabla 2

Visualizaciones

Matriz de Correlaciones

La matriz de referencia muestra los coeficientes de correlación entre las distintas variables numéricas del dataset. Puede tomar los siguientes valores:

- valor 1, representa una correlación positiva perfecta
- valor -1, representa una correlación negativa perfecta
- valor 0, no existe correlación lineal
- valores cercanos a 0, representan una relación débil
- valores cercanos a 1 o -1, representan una relación fuerte.

Analizando algunos valores para facilitar la comprensión de la misma, podemos mencionar los siguientes casos:

- Annual_Income tiene una correlación de 0.783 con Monthly_Balance, lo que indica una relación positiva fuerte, al aumentar el primero también aumenta el segundo.
- Outstanding_Debt tiene una correlación negativa de -0.629 con Credit_History, lo que implica que mientras más alta es la deuda pendiente, menos antigüedad tiene su historial crediticio.
- Credit_Utilization_Ratio y Age tienen una correlación de -0.005, lo que implica que prácticamente no existe relación lineal entre ambas variables.

El gráfico 1 contiene la matriz completa para poder visualizar todas las correlaciones y entender las relaciones entre las distintas variables.

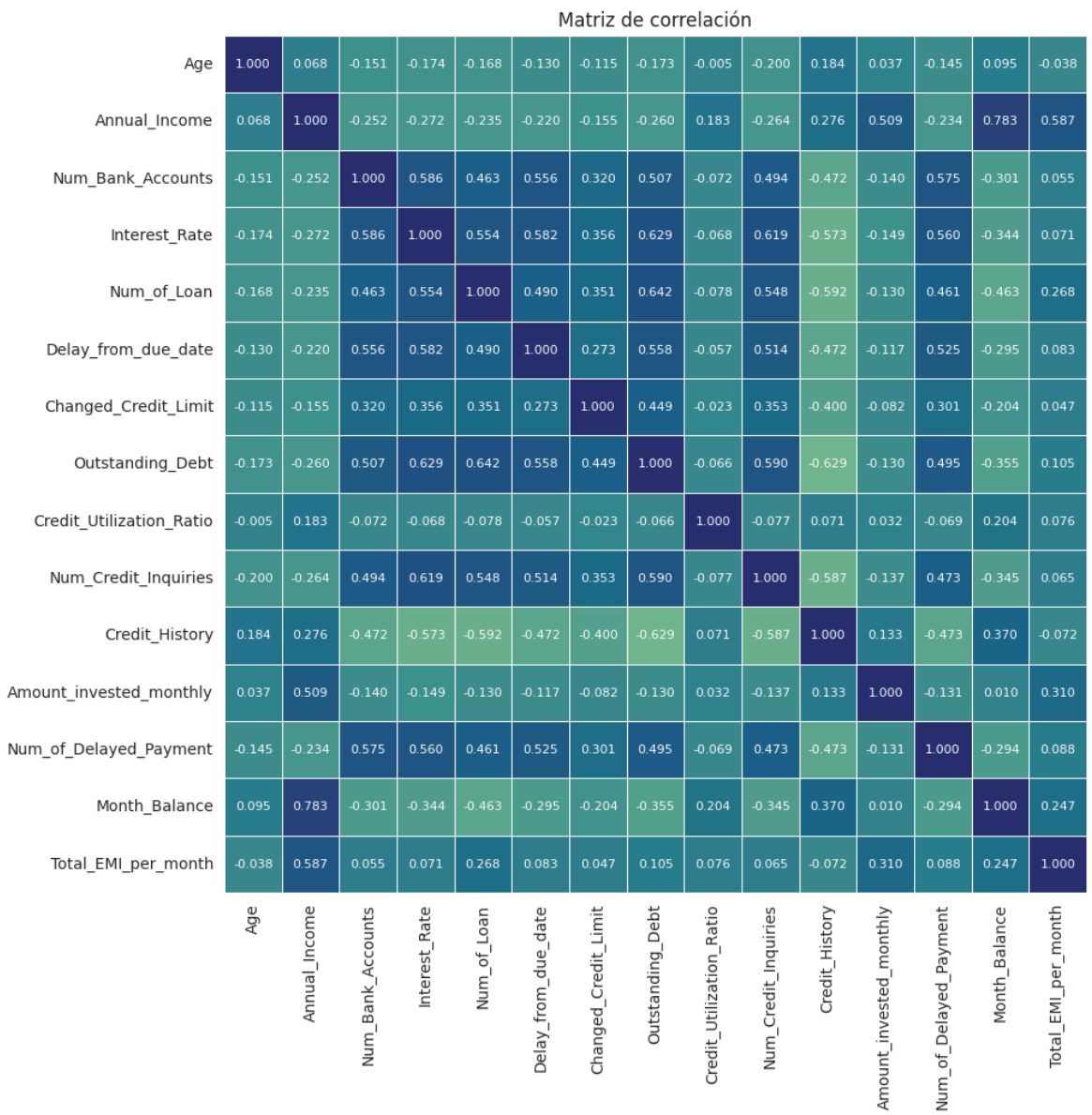

Gráfico 1

Categorías / Proporciones

Credit Score

Existen tres categorías para nuestra variable target:

- Good (buena)
- Standard (media)
- Poor (mala)

El gráfico 2.a permite visualizar las cantidades, y el gráfico 2.b las proporciones de cada categoría.

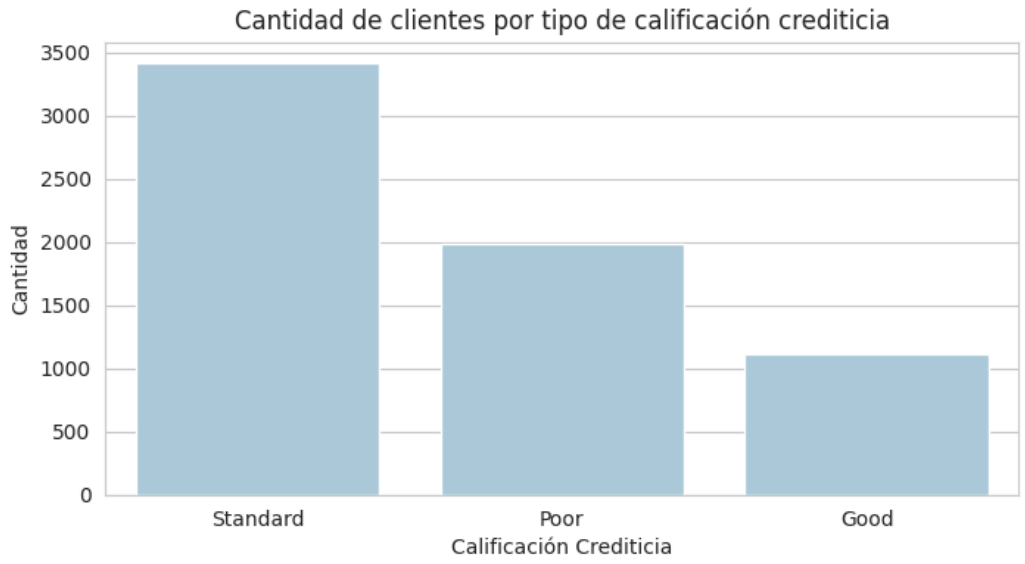

Gráfico 2.a

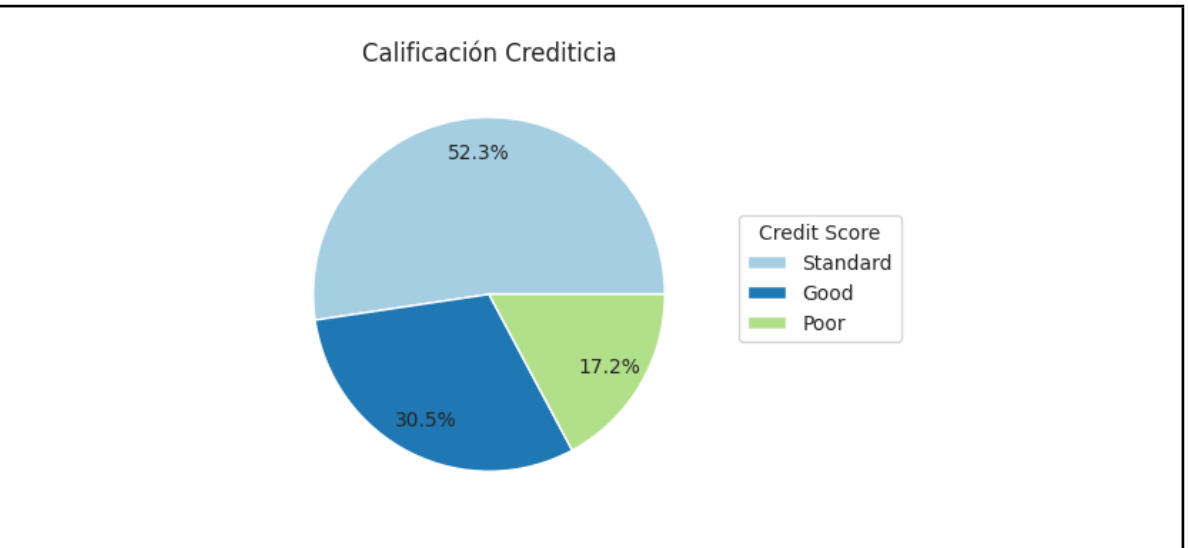

Gráfico 2.b

Payment Behavior

Para esta variable existen las siguientes categorías:

- High spent - Medium value payments, que categoriza a los clientes de alto gasto y pago medio de su saldo pendiente
- Low spent - Small value payments, categoriza a los clientes de gasto bajo y pagos pequeños de su saldo pendiente (pago mínimo, probablemente).
- High spent - Large value payments, categoriza a los clientes de alto gasto y realiza pagos grandes.
- Low spent - Medium value payments, categoriza a los clientes de bajo gasto y pagos medios.
- Low spent - Large value payments, para categorizar a los clientes de bajo gasto y pagos grandes.
- High spent - Small value payments, categoriza clientes de alto gasto y pagos pequeños, lo que podría implicar una señal de riesgo ya que su deuda aumenta rápidamente.

Las principales categorías son la de High spent - Medium value payments y Low spent - Small value payments, que representan el 47,7% de los casos según lo que se puede ver en el gráfico 3.

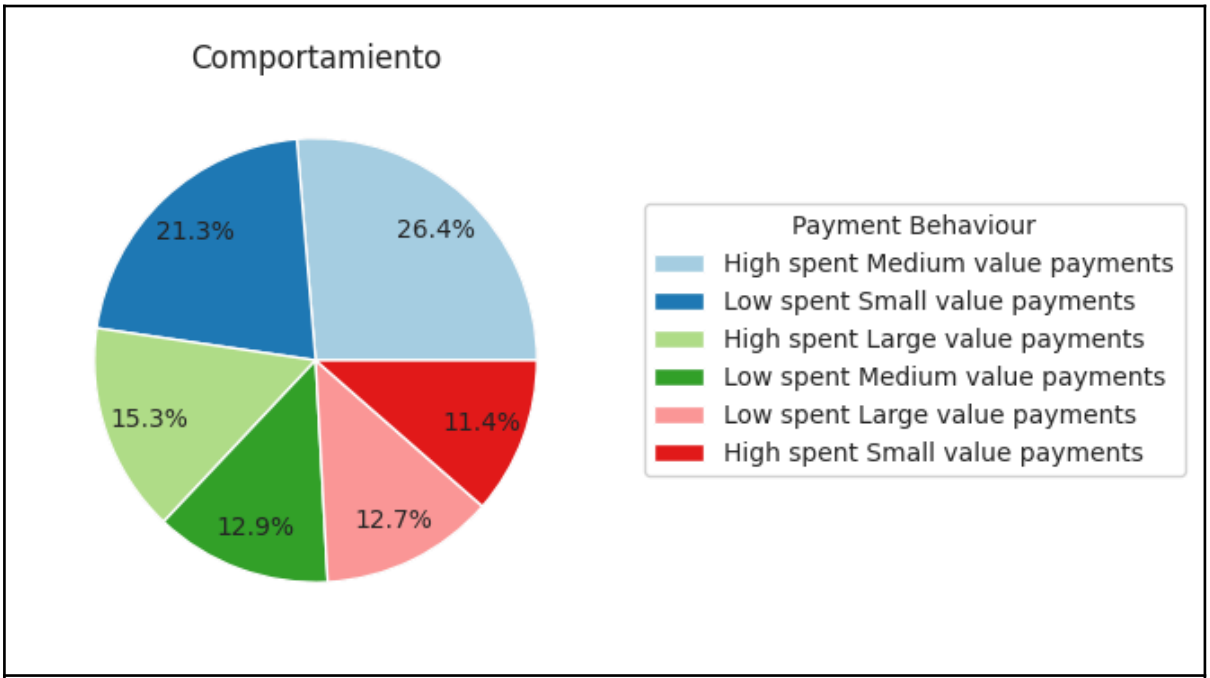

Gráfico 3

Distribuciones

Annual Income

Para analizar esta variable uso como referencia el gráfico 4 y la tabla 3.
Esta variable tiene una asimetría hacia la derecha para cada categoría. En el caso de la categoría *Poor* esto se ve más marcado ya que el primer cuartil, la mediana y media presentan los valores más bajos, así como menos desvío. En el caso de la categoría *Good*, los valores para las medidas mencionadas anteriormente presentan los valores más altos, así como una curva más plana dada la menor cantidad de casos.
Todas las categorías presentan valores mínimos y máximos similares para esta variable.

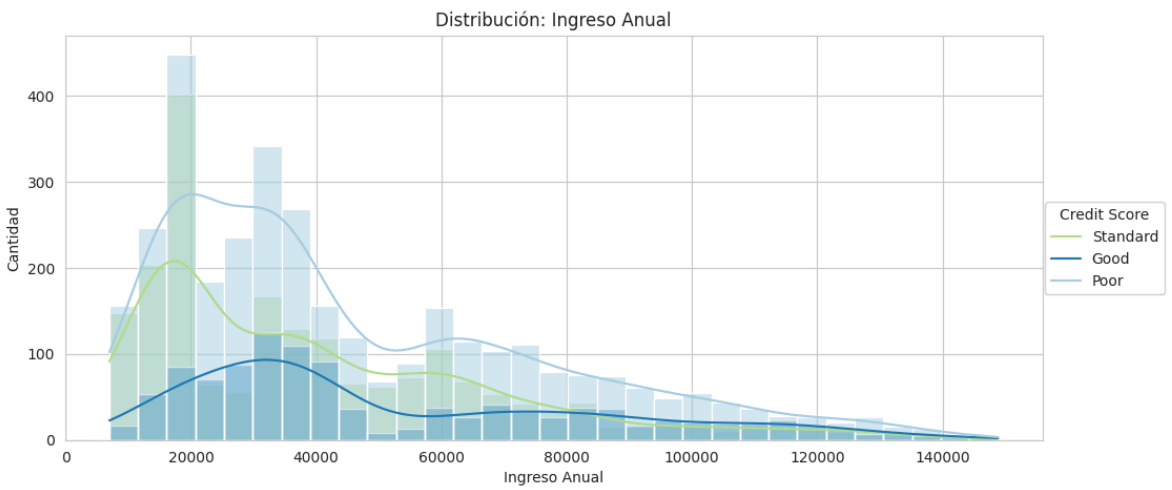

Gráfico 4

	Annual_Income							
	count	mean	std	min	25%	50%	75%	max
Credit_Score								
Good	1118.0	52703.655291	33022.500968	7189.54	28717.0175	39536.035	75871.0375	148786.04
Poor	1989.0	40340.790762	28272.024038	7012.31	17756.4200	33271.740	57439.1200	146771.72
Standard	3411.0	46333.158065	30712.839444	7006.52	20786.5500	35579.680	65175.1700	148841.56

Tabla 3

Interest Rate

Esta variable presenta distintas distribuciones según el credit score. Para las calificaciones *Good* y Poor encuentro asimetría hacia la derecha, y para la categoría restante, hacia la izquierda, según lo expuesto en el gráfico 5. El valor correspondiente al tercer cuartil para la categoría *Good* es menor que el valor del primer cuartil de la categoría *Poor*.
Las medias y las medianas para cada calificación crediticia tienen valores cercanos.

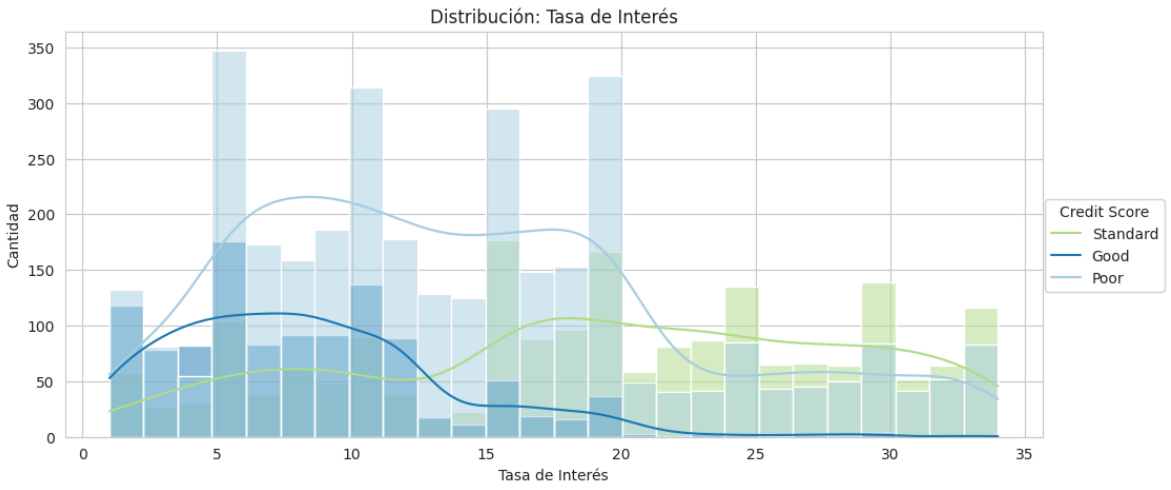

Gráfico 5

	Interest_Rate							
	count	mean	std	min	25%	50%	75%	max
Credit_Score								
Good	1118.0	8.370304	5.176793	1.0	5.0	8.0	11.0	33.0
Poor	1989.0	19.320764	8.913381	1.0	13.0	20.0	27.0	34.0
Standard	3411.0	14.235122	8.044514	1.0	8.0	13.0	19.0	34.0

Tabla 4

Monthly Balance

Esta última variable estudiada presenta asimetría hacia la derecha para las tres categorías de credit score.

En el caso de la categoría Good el valor de la media es mayor a la mediana, lo que implica una cola con mayor probabilidad de encontrar valores superiores a la media en relación a las distribuciones de las otras categorías de calificación crediticia.

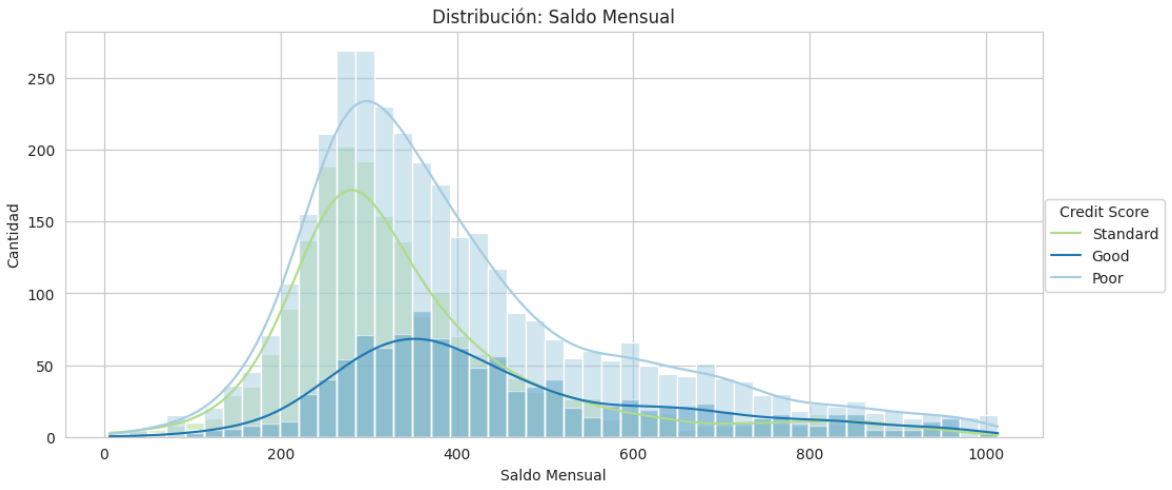

Gráfico 6

	Month_Balance							
	count	mean	std	min	25%	50%	75%	max
Credit_Score								
Good	1118.0	454.412293	188.614252	36.997273	322.369127	399.184717	567.120063	1013.505949
Poor	1989.0	353.574124	164.629574	5.958076	253.597403	309.017071	402.847995	999.674319
Standard	3411.0	406.662443	186.195553	6.241510	277.358956	353.702334	492.472089	1012.552737

Tabla 5

Scatterplot

Para el análisis de los siguientes gráficos se recomienda consultar la matriz de correlaciones (gráfico 1).
En la sección correspondiente en la notebook se pueden encontrar otros gráficos correspondientes a otras relaciones entre variables además de las expuestas en este apartado.

Annual Income vs Monthly Balance

Estas variables presentan una correlación positiva de 0.783, lo que implica una relación lineal fuerte entre ambas.
El gráfico 7 deja en evidencia la misma, con mayor concentración de valores bajos tanto para ingreso anual como para saldo mensual.
La relación implica que a mayor ingreso anual, mayor es el saldo mensual del cliente.

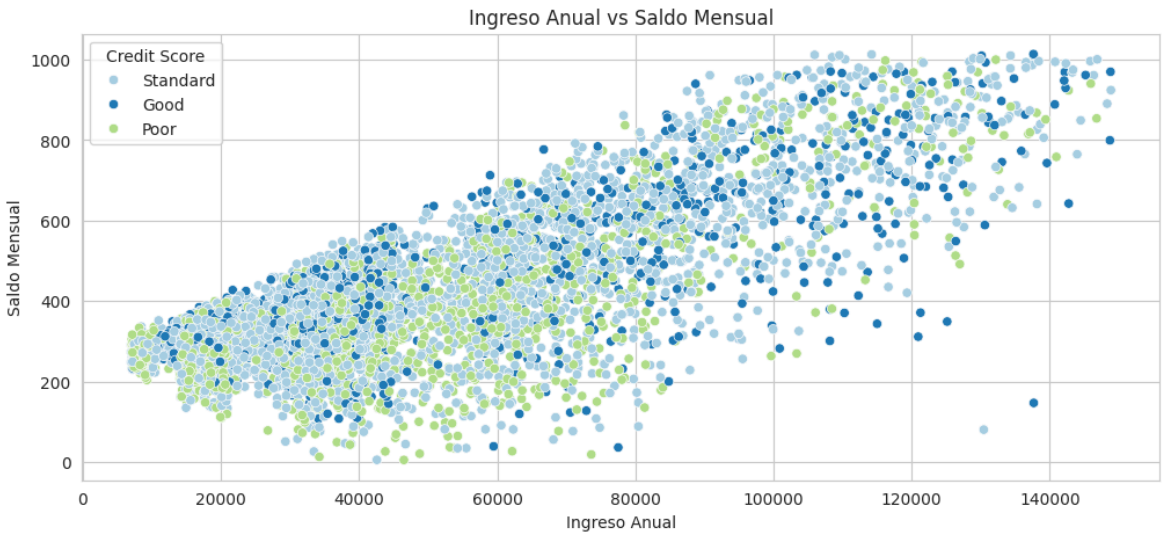

Gráfico 7

Outstanding Debt vs Credit History

La correlación entres estas variables es negativa de -0.629, lo que implica una alta relación lineal entre ambas.
El gráfico 8 muestra esa relación en la cual a mayor deuda pendiente, menor antigüedad de historia crediticia.

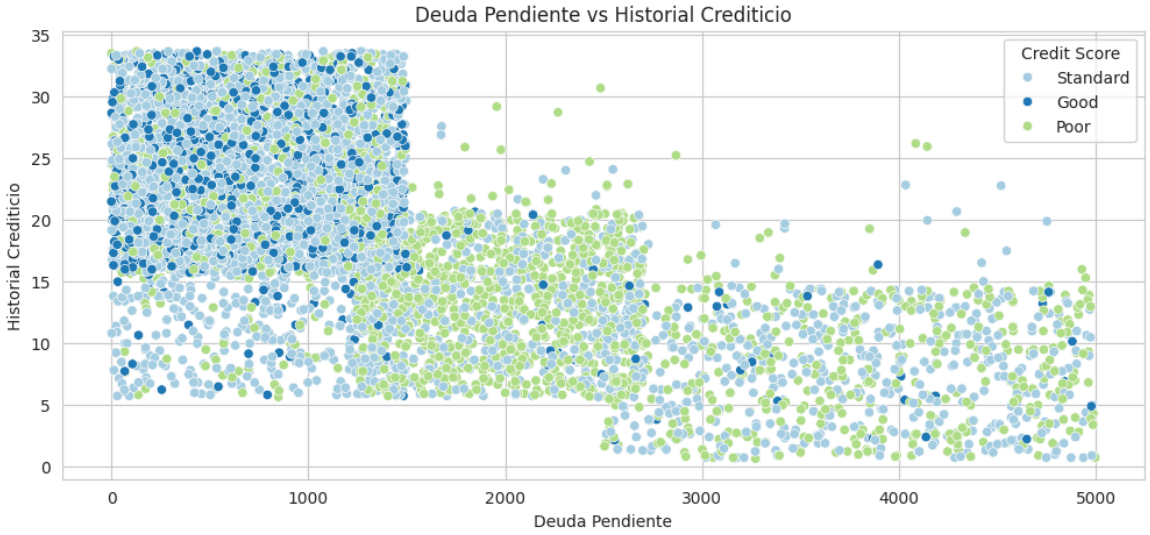

Gráfico 8

Age vs Credit Utilization Ratio

La correlación entre estas variables es de -0.005, lo que implica que la relación lineal entre ambas es nula. Al ver el gráfico 9 queda en evidencia que no se distingue ninguna relación.

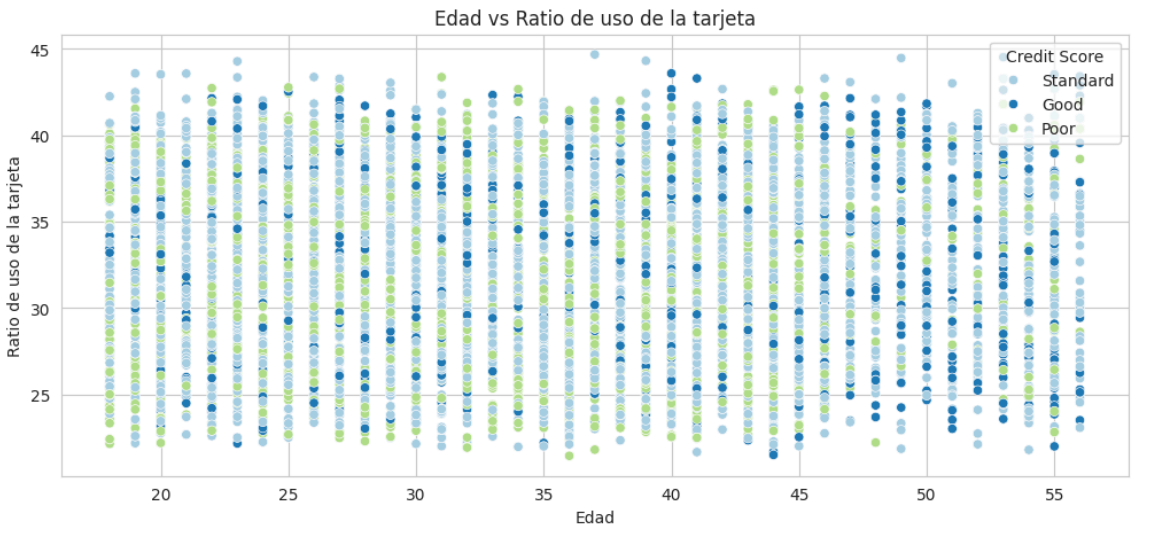

Gráfico 9

Conclusiones

Tanto en la notebook como en el presente documento no se incluyó el gráfico *pairplot* ya que el mismo tiene una demora notable al ejecutarse abierto por credit score. Las variables numéricas no siguen una distribución normal, la mayoría presentan asimetría a la derecha.
Según lo expuesto en la matriz de correlaciones existen un grupo de variables que tienen correlación positiva media a fuerte como ser la que existe entre las siguientes variables: cantidad de cuentas bancarias, tasa de interés, cantidad de préstamos, demora desde fecha de pago, cambio en límite crediticio y deuda pendiente.
En el caso de la antigüedad del historial crediticio tiene una relación lineal negativa media con cantidad de cuentas bancarias, tasa de interés, cantidad de préstamos, demora desde fecha de pago, cambio en límite crediticio y deuda pendiente. Estas mismas variables tienen relación lineal positiva media y media-fuerte con cantidad de pagos atrasados.
La relación más fuerte existe entre el ingreso anual y el saldo mensual.
En la próxima etapa del proyecto, antes de seleccionar los algoritmos para desarrollar el modelo de clasificación de la calificación crediticia corresponde continuar con el preprocesado de los datos. Queda pendiente para esa etapa del análisis de las variables con varianza cero o cercano a cero para definir si conviene incluirlas o no en el modelo final.
Dada la asimetría de las distribuciones, no se descarta la normalización o estandarización de las mismas para trabajar con distribuciones normales, para evitar el efecto de aquellas variables que tengan mayor varianza y puedan afectar al resultado de la clasificación.
Por último, queda la binarización de las variables categóricas para poder luego empezar a trabajar con los algoritmos.