Proyecto Final

Título:

Modelo supervisado de clasificación para clientes de un banco

Valeria Pérez Cong Sánchez

CU 145009

ITAM

Aprendizaje de Máquina

Fernando Esponda Darlington

ÍNDICE

INTRODUCCIÓN

Ray Kurzweil, Director de Ingeniería en Google y encargado de dirigir el desarrollo de Inteligencia Artificial, se ha dedicado a hacer predicciones sobre el surgimiento de nuevas tecnologías y su respectivo impacto. Una de sus predicciones más impactantes es que para el año 2029 la inteligencia artificial va a alcanzar niveles humanos de inteligencia. El hecho de que "las máquinas" nos ayuden a analizar información representa ventajas importantes como poder analizar grandes volúmenes de información en poco tiempo y la posibilidad de tomar decisiones objetivas basadas en datos y no solo basadas en la intuición o sentimientos.

Un caso muy conocido y relevante es poder clasificar si un cliente va a poder a pagar o no su tarjeta de crédito según su información demográfica y su historial de pagos. Esta información es muy valiosa para los bancos pues así pueden saber si seguir otorgando un crédito a los clientes o no hacerlo. Estar financiando a un cliente que no es capaz de pagar el crédito otorgado puede representar grandes pérdidas para los bancos.

El set de datos se obtuvo de Kaggle.com y fue provisto por UCI Machine Learning. El set de datos contiene información de pagos que cayeron en default (no se pagaron), factores demográficos, datos del crédito historial de pagos, cantidad a pagar y pagos realizados por clientes de tarjeta de crédito de un banco en Taiwán de Abril 2005 a Septiembre 2005. El set de datos contiene 30,000 observaciones y 25 variables.

En este estudio se busca encontrar el o los modelos de clasificación adecuados para poder saber si un cierto cliente podrá pagar su tarjeta de crédito o no según ciertos parámetros. Para seleccionar el mejor modelo, se utilizará la curva ROC como medida de desempeño, así como las medidas de Accuracy, Precision y Recall. Es decir, proporción de clasificaciones correctas e incorrectas según el valor "deseado" de la variable target.

METODOLOGÍA

Análisis Exploratorio De Los Datos

Antes de aplicar los modelos, se hizo el análisis exploratorio de los datos para saber cuáles eran las variables del set, su cardinalidad, valores nulos y relación entre variables.

Tamaño del set de datos

Como se mencionó anteriormente, el set de datos cuenta con 30,000 observaciones y 24 variables. Originalmente, el set contaba con 25 variables pero una de ellas era la variable ID, que es un valor único de identificación para cada cliente y que no aparta información, por lo que decidimos descartarla.

Variables en el set y su clase

Variable	Clase	Variable	Clase
LIMIT_BAL	numeric	BILL_AMT2	integer
SEX	integer	BILL_AMT3	integer
EDUCATION	integer	BILL_AMT4	integer
MARRIAGE	integer	BILL_AMT5	integer
AGE	integer	BILL_AMT6	integer
PAY_0	integer	PAY_AMT1	integer
PAY_2	integer	PAY_AMT2	integer
PAY_3	integer	PAY_AMT3	integer
PAY_4	integer	PAY_AMT4	integer
PAY_5	integer	PAY_AMT5	integer
PAY_6	integer	PAY_AMT6	integer
BILL_AMT1	integer	default	integer

A continuación se explican las variables:

- LIMIT_BAL : es la cantidad de crédito otorgado a cada cliente. Se puede observar que en el set hay créditos desde 10,000 dólares (493 clientes) hasta 1’000,000 (1 solo cliente).

- SEX: género del cliente. El 1 representa hombres y el 2 representa mujeres. Se puede observar que hay más mujeres que hombres en el set. Variable categórica.

- EDUCATION: el nivel de educación de cada cliente. Donde cada número representa una categoría diferente: el número 1 es graduate school; el número 2 es university; el número 3 es high school; el número 4 es otros; números 5 y 6 son desconocidos. Afortunadamente, pocas observaciones pertenecen a las categorías 4,5 y 6. En la página de UCI, de donde se obtuvo el set de dato, no se específica el significado de la categoría 0; se va a suponer que el 0 implica ausencia de atributo, o sea, gente con escolaridad menor a graduate school.

- MARRIAGE: estado civil del cliente. Variable categórica donde 0 representa otros, 1 representa casado, 2 representa soltero y 3 divorciado .

- AGE :edad del cliente, donde podemos ver que hay clientes desde los 21 años hasta los 79 años.

- PAY_NUM: es una variable categórica para saber el historial de pagos para meses pasados, cada número representa un estado de pago diferente: -2 significa que no consumió el crédito; -1 significa que pagó a tiempo y la deuda completa; 0 uso de crédito renovable; 1 significa que retrasó el pago por un mes; 2 significa que atrasó el pago por 2 meses… 9 significa que atrasó el pago por 9 meses o más.

- BILL_AMTNUM: Variable contínua del monto a pagar cada mes. Donde BILL_AMT1 es el monto a pagar para septiembre del 2005 y BILL_AMT2 es el monto a pagar en agosto 2005. Número negativos implican un saldo a favor del cliente.

- PAY_AMTNUM: Monto pagado cada mes. Siguiendo la misma lógica que en la variable anterior.

- Default: variable categórica donde 0 implica que no cayó en default (sí pudo pagar) y 1 implica que sí cayó en default (no pudo pagar). El 77.88% de los clientes en el set no cayeron en default, mientras que el 22.12% sí cayeron.

De las 24 variables presentes en el set de datos hay 10 categóricas: SEX, EDUCATION, MARRIAGE, PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6 y la variable default. Hacemos la distinción pues el análisis exploratorio de las variables categóricas y numéricas continuas es distinto.

Valores Nulos

Variable	Valores Nulos	Variable	Valores Nulos
LIMIT_BAL	0	BILL_AMT2	3
SEX	0	BILL_AMT3	0
EDUCATION	0	BILL_AMT4	5
MARRIAGE	0	BILL_AMT5	8
AGE	0	BILL_AMT6	5
PAY_0	0	PAY_AMT1	16
PAY_2	0	PAY_AMT2	18
PAY_3	0	PAY_AMT3	9
PAY_4	0	PAY_AMT4	0
PAY_5	0	PAY_AMT5	0
PAY_6	0	PAY_AMT6	12
BILL_AMT1	2	default	0

Se puede observar que la variable BILL_AMT1 cuenta con 2 valores nulos o vacíos, BILL_AMT2 cuenta con 3, BILL_AMT4 cuenta con 5, BILL_AMT5 cuenta con 8, BILL_AMT6 cuenta con 5, PAY_AMT1 cuenta con 16, PAY_AMT2 cuenta con 18, PAY_AMT3 cuenta con 9 y PAY_AMT6 cuenta con 12 valores nulos o vacíos. Más adelante se va a definir el tratamiento que se les dará a los valores vacíos.

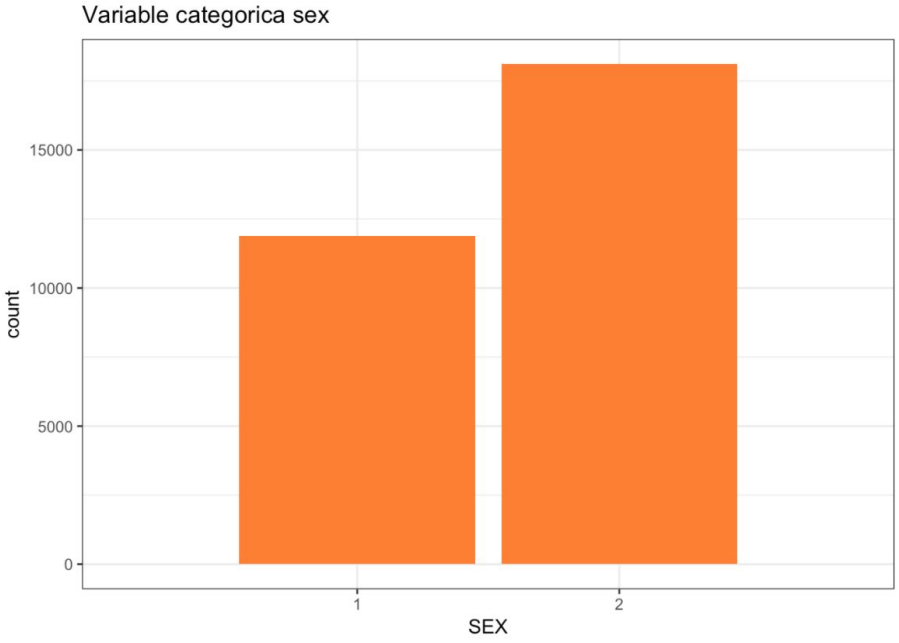
Cardinalidad de las variables

Variable	Cardinalidad	Variable	Cardinalidad
LIMIT_BAL	81	BILL_AMT2	22345
SEX	2	BILL_AMT3	22026
EDUCATION	7	BILL_AMT4	21546
MARRIAGE	4	BILL_AMT5	21007
AGE	56	BILL_AMT6	20604
PAY_0	11	PAY_AMT1	7942

PAY_2	11	PAY_AMT2	7897
PAY_3	11	PAY_AMT3	7518
PAY_4	11	PAY_AMT4	6937
PAY_5	10	PAY_AMT5	6897
PAY_6	10	PAY_AMT6	6938
BILL_AMT1	2273	default	2

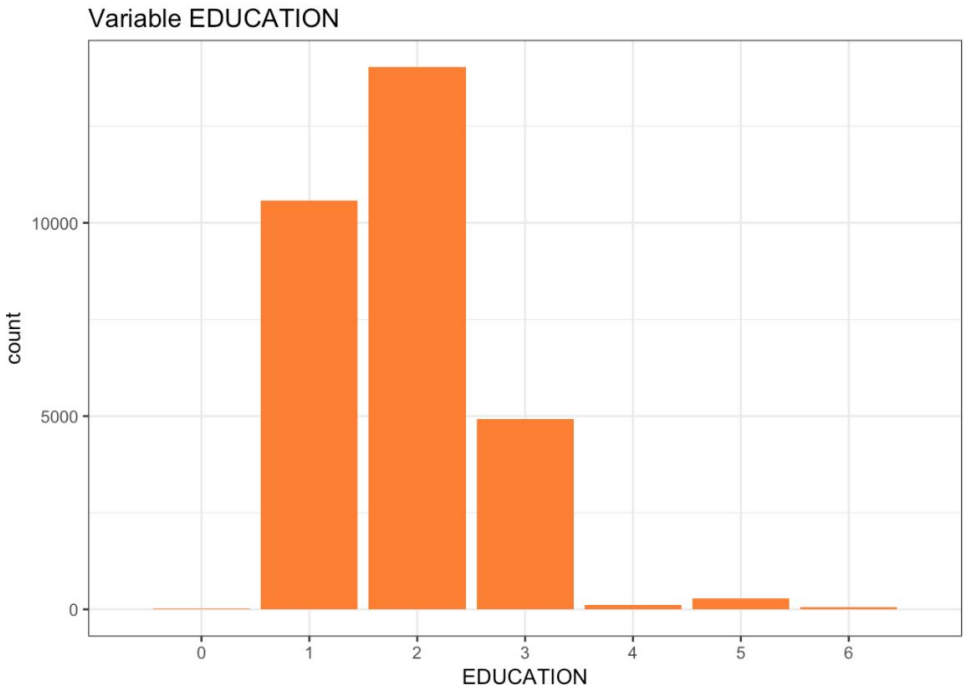
Naturalmente, se puede observar que las variables categóricas son las que cuentan con cardinalidad menores o iguales a 11. Por otro lado, las variables numéricas continuas cuentan con cardinalidades muy altas.

Variable categórica sex



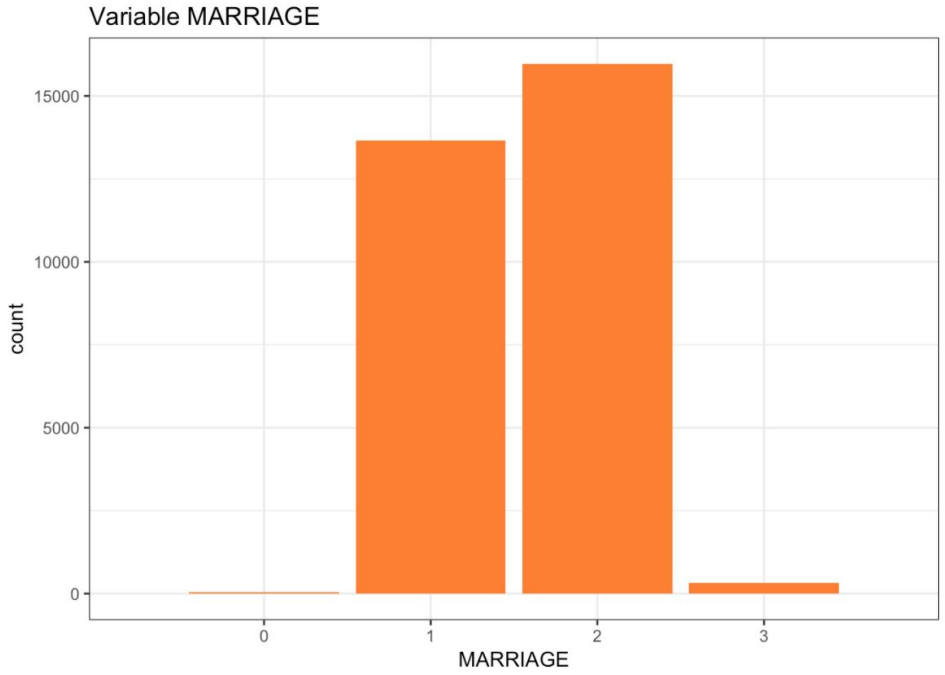
Se puede observar que el set de datos está desproporcionado por esta categoría pues la mayoría de las observaciones pertenecen a la categoría 2 (female).

Variable categórica education



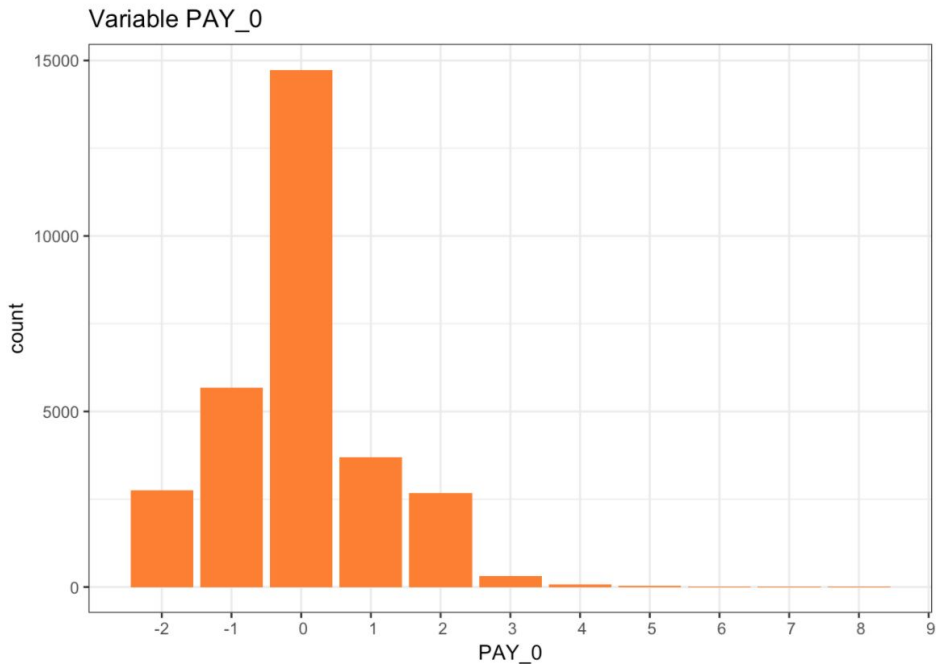
Muy pocas observaciones pertenecen a las categorías 0,4,5,6 que representan niveles de educación desconocidos. La categoría más popular es la 2 (nivel universitario), le sigue la categoría 1 (nivel graduate school) y después la categoría 3 (nivel preparatoria).

Variable categórica marriage



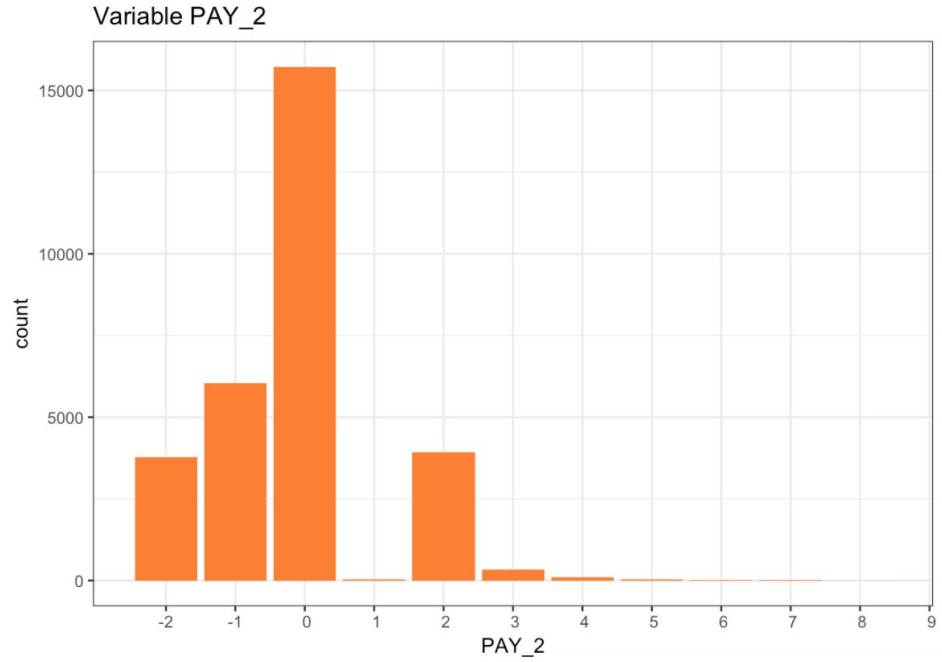
Se puede observar que la mayoría de los clientes en el set pertenecen a la categoría 2 (solteros), le sigue la categoría 1 (casados) y muy pocos pertenecen a la categoría de divorciados u otros estados civiles (valores 0 y 3).

Variable categórica Pay_0



Para el status del pago de Septiembre 2005, la mayoría de los clientes uso crédito renovable (valor 0); después, la segunda mayoría de los clientes pagaron a tiempo (valor -1); la tercera categoría más popular fue para lo clientes que difirieron el pago un mes (valor 1); el número clientes que no usaron el crédito (-2) y los que difirieron el pago 2 meses (valor 2) es muy similar. Pocos clientes difirieron el pago por 3 meses o más.

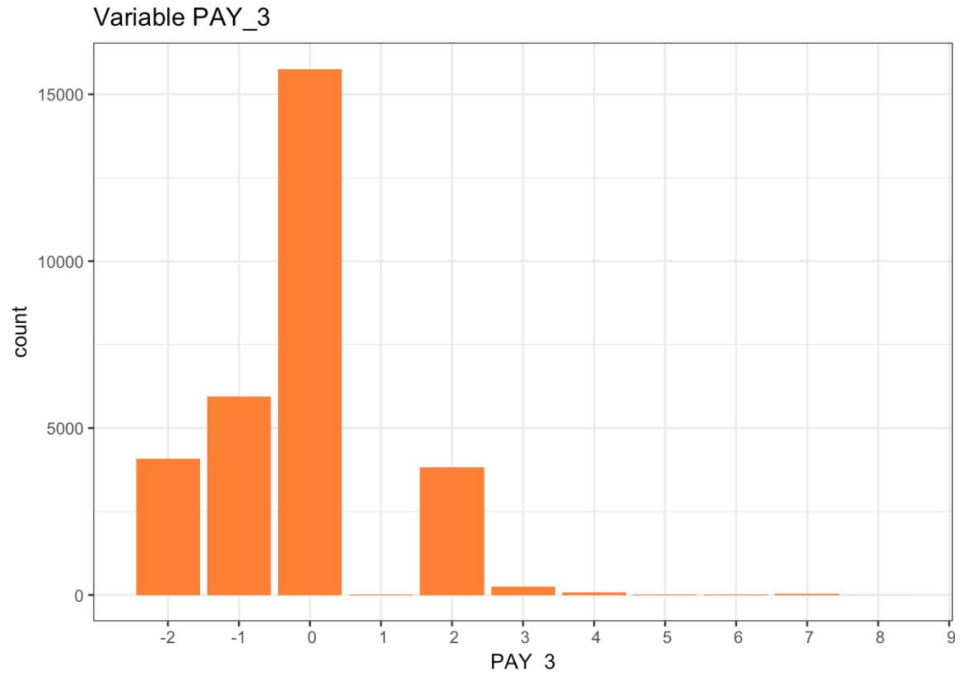
Variable categórica Pay_2



Para esta variable vemos un comportamiento distinto al de la variable Pay_0. Esta variable corresponde al status del pago para Agosto 2005. Nuevamente, vemos que más de 15 mil clientes hicieron uso del crédito renovable (categoría 0), la segunda categoría con más

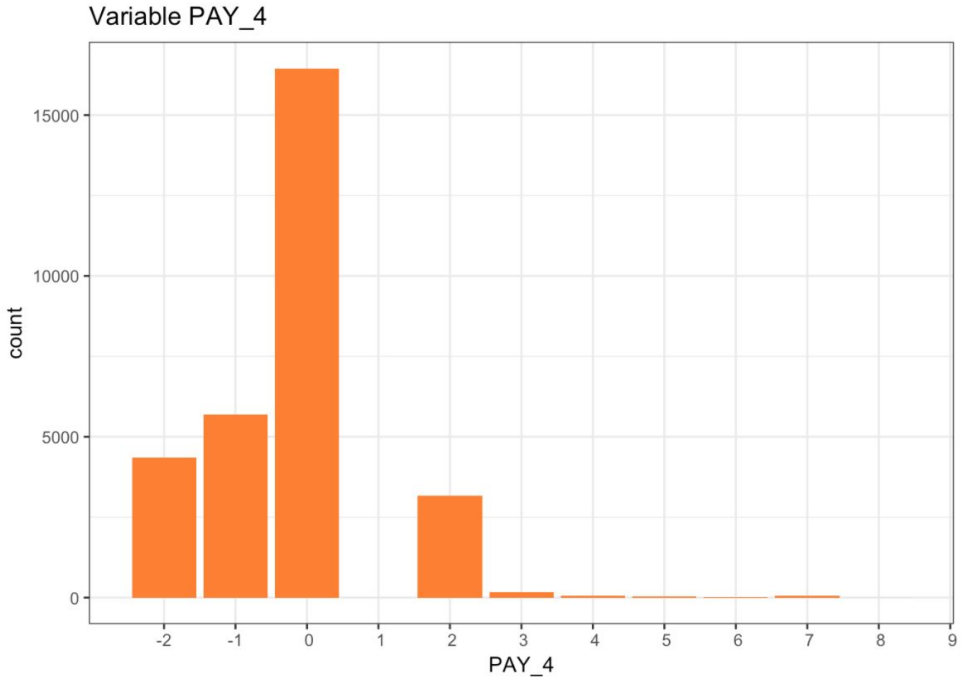
clientes es la de los clientes que sí pagaron a tiempo (valor -1). Muy pocos clientes difirieron sus pagos por 1, 3 o más meses. Cerca de 3750 clientes difirieron sus pagos por 2 meses.

Variable categórica Pay_3



Para el estado de los pagos de Julio 2005, se puede observar un comportamiento muy parecido al de la variable Pay_2 (pagos para Agosto 2005).

Variable Categórica Pay_4

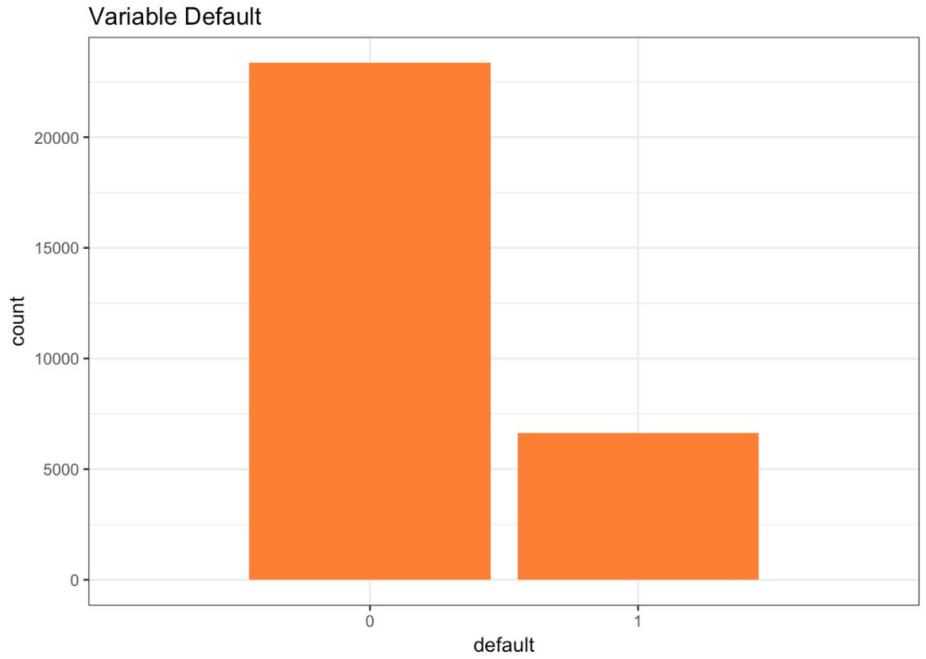


Nuevamente, para el estado de los pagos de Junio 2005 se puede observar un comportamiento muy similar al de la variable Pay_2 (Agosto 2005) y Pay_3 (Julio 2005). El comportamiento

de los datos que se observa para las variables Pay_5 y Pay_6 es muy parecida a la variable Pay_4 por lo que no se incluyen esas gráficas de barras.

En general, se observa que el valor que predomina es el cero (clientes que hicieron uso del crédito renovable), el segundo valor con más observaciones es el -1 (clientes que pagaron a tiempo), el tercer valor con más observaciones es el -2 (personas que no usaron el crédito). Cerca de 2500 clientes difirieron el pago de la tarjeta por 2 meses, muy pocos lo difieron por 3 meses o más.

Variable categórica default (la variable target)



Se puede observar que la mayoría de los clientes en el set no cayeron en default (valor). Esto quiere decir que el set de datos no está balanceado según la variable target. Es posible que este desbalance traiga problemas al momento de aplicar los modelos de clasificación.

Resumen Variables Numéricas

Variable	Cardinalidad	Nulos	Mínimo	Máximo	Media	Desviación	Primer Cuartil	Mediana	Tercer Cuartil
LIMIT_BAL	81	0	10000	1000000	167484.	129747.66	50000	140000.0	240000.00
AGE	56	0	21	79	35.49	9.217904	28	34.0	41.00
BILL_AMT1	22753	2	-165580	964511	51220.07	73637.23	3558.25	22379.0	67087.75
BILL_AMT2	22345	3	-69777	983931	49163.99	73637.23	2984.00	21197.0	63995.00
BILL_AMT3	22026	0	-157264	1664089	47013.15	69349.38	2666.25	20088.5	60164.75

BILL_AMT4	21546	5	-170000	891586	43243.48	64312.26	2323.00	19048.0	54470.00
BILL_AMT5	21007	8	-81334	927171	40272.14	60733.35	1761.75	18097.0	50149.75
BILL_AMT6	20604	5	-339603	961664	38861.57	59553.84	1256.00	17067.0	49146.50
PAY_AMT1	7942	16	0	873552	5606.57	16344.53	1000.00	2100.0	5006.00
PAY_AMT2	7897	18	0	1684259	5851.34	22833.52	832.00	2008.0	5000.00
PAY_AMT3	7518	9	0	896040	5197.24	17532.87	390.00	1800.0	4500.00
PAY_AMT4	6937	0	0	621000	4826.07	15666.15	296.00	1500.0	4013.25
PAY_AMT5	6897	0	0	426529	4799.38	15278.30	252.50	1500.0	4031.50
PAY_AMT6	6938	12	0	528666	5167.56	17597.45	116.00	1500.0	4000.00

Se puede observar que hay 14 variables continuas. A continuación un análisis de las variables:

- La variable de LIMIT_BAL cuenta con 81 valores diferentes, 0 valores nulos, su valor mínimo es 10,000, su valor máximo es de 1’000,000, la media de los créditos otorgados es de 167,484.3227 y la desviación estándar es de 129,747.661567 lo cual resulta bastante alto. El 25% de los datos tiene un límite de crédito de 50,000 o menos, el 50% de los clientes cuenta con un límite de crédito de 140,000 o menos, y un 25% tiene un límite de 240,000 o más.

- La variable de AGE cuenta con 56 valores diferentes, 0 valores nulos, una edad mínima de 21 años y una máxima de 79, la edad media es de 35.5 años aproximadamente, hay una desviación estándar de 9.2 años. El 25% de los clientes tiene 28 años o menos, el 50% tiene 34 años o menos y el 25% tiene 41 años o más.

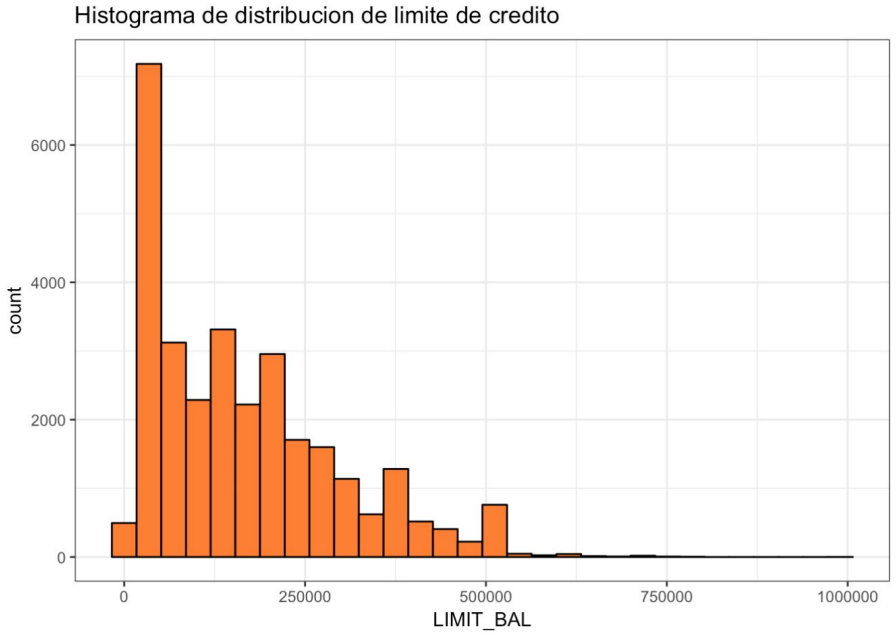
- La variable BILL_AMOUNT1 tiene 22723 valores diferentes (claramente es una variable contínua), 2 valores nulos, el valor mínimo es de -165580 (saldo a favor), el valor máximo del monto a pagar es de 964,511 dólares, el saldo a pagar promedio por los clientes para este mes es de 51,220 dólares, con una desviación estándar de 73,637.2 dólares lo cual es sumamente alto e incluso mayor a la media. El 25% de los clientes tienen un saldo a pagar de 3558 o menos (incluyendo saldo a favor), el 50% de los clientes tiene saldo de 22,379.0 o menos, y el 25% de los clientes tiene un saldo de 67,087.75 o más. En las demás variables de BILL_AMOUNT se observa un comportamiento similar, donde hay clientes con saldos a favor, saldos a pagar máximos cercanos a 90,000, una media cercana o mayor a 40,000, y una desviación estándar sumamente alta de 60,000 o más y que es mucho mayor a la media.

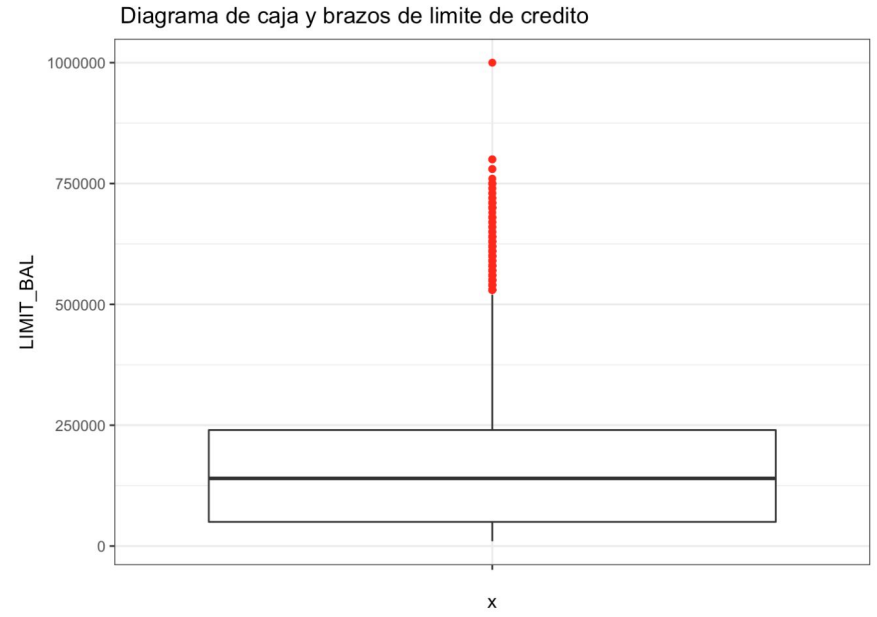
- La variable PAY_AMT1 tiene una cardinalidad de 7942, 16 valores nulos, un valor mínimo de 0 dólares, un valor máximo de 873,552 dólares pagados, la media de pagos

es de 5,606.6 dólares, la desviación estándar es de 16,344.5 lo cual es bastante alto y casi 3 veces mayor a la media. El 25% de los pagos fue de 1000 dólares o menos, el 50% de los pagos fueron de 2100 dólares o menos, y el 25% de los pagos fueron de 5006.00 o más. Se puede observar que la media de los datos está muy cerca del tercer cuartil. En las demás variables de este tipo (PAY_AMOUNT) también se observa una media mucho menor a la desviación estándar y la media contenida en el tercer cuartil.

En general, se puede observar que el rango de las variables es muy alto, lo que provoca que la desviación estándar sea muy alta e incluso mayor que la media.

Variable Numérica Limit_Bal

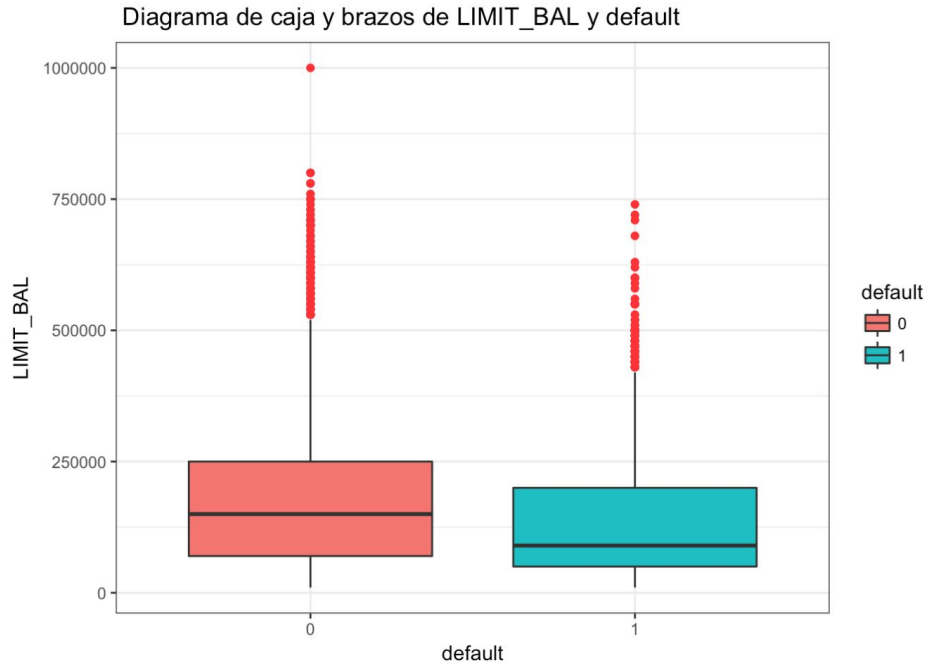




Se puede observar que la distribución de la variable Limit_Bal está concentrada entre 0 y 250,000. La mediana de los datos está en 140,000. Se puede observar que los valores outliers

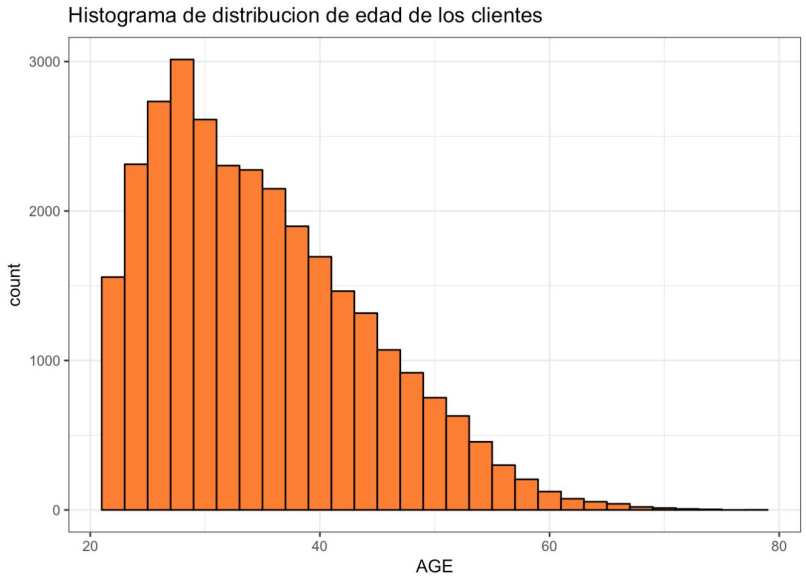
empiezan a partir de 530,000. En el set hay 167 observaciones de esta variable con valores outliers.

Relación con la variable target:

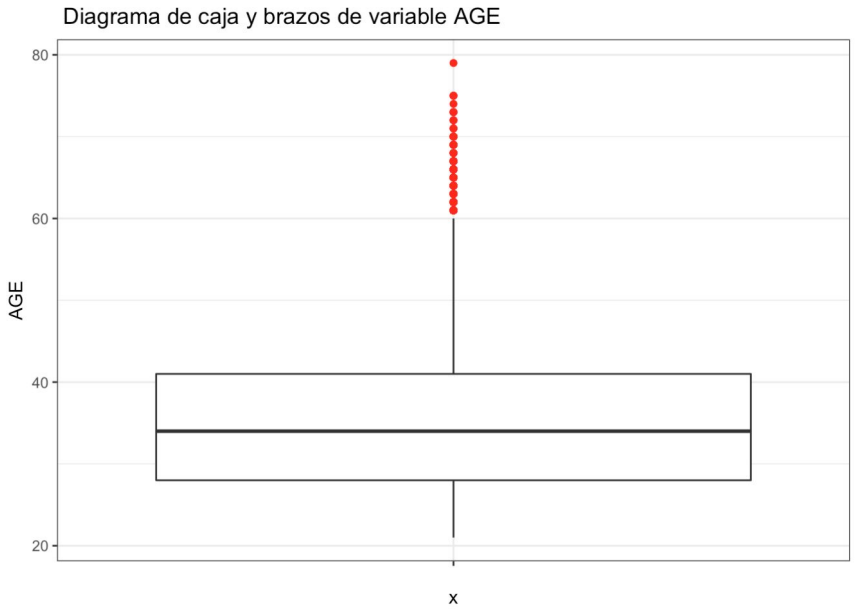


Se puede observar que la distribución de la variable de límite de crédito es diferente para los clientes que no cayeron en default (color rosa) y para los clientes que sí cayeron en default. La mediana del límite de crédito para los clientes que no cayeron en default es mayor que para los clientes que sí cayeron. En general, los cuartiles son más altos para los clientes que no cayeron en default. Además, límites de crédito que ya forman parte de los outliers para el valor de 1 son observaciones no atípicas dentro de la distribución de los clientes que no cayeron en default.

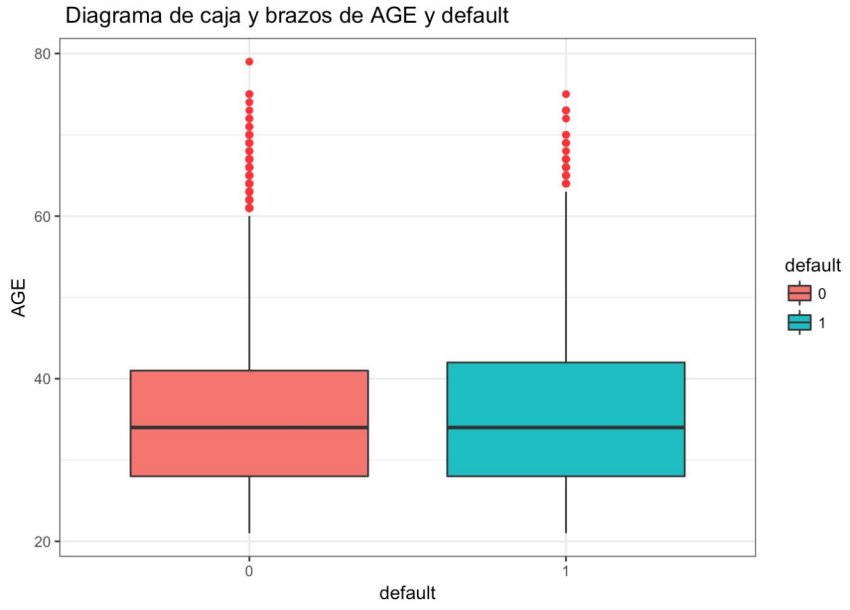
<u>Variable numérica Age</u>



La distribución de las edades para los clientes en el set tiene una forma parecida a una distribución ji-cuadrada, pues muestra un sesgo a la derecha. Los datos están concentrados entre los 20 y 40 años, lo cual hace sentido pues 41 años es el tercer cuartil.



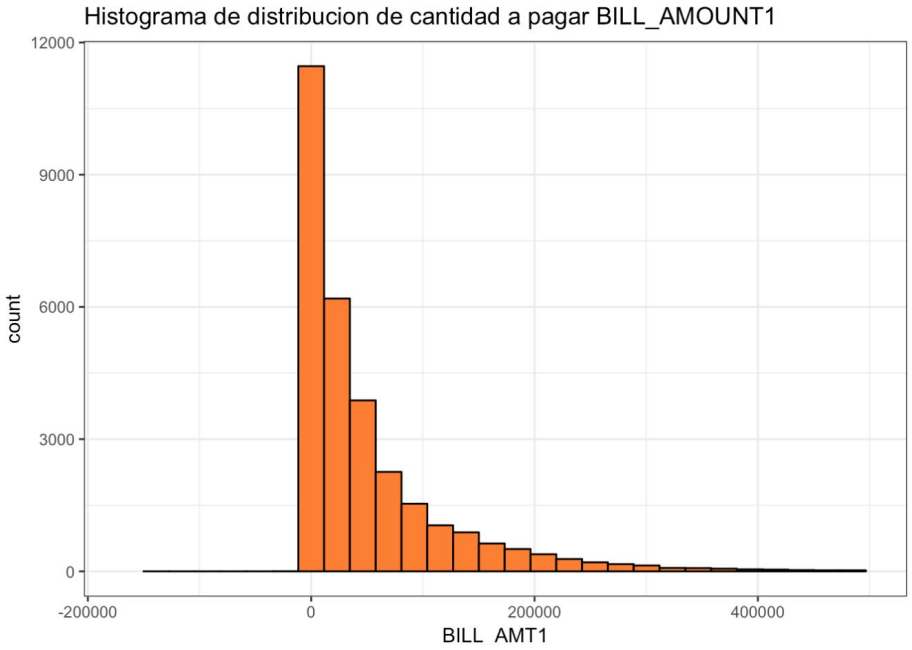
Se puede observar que el primer cuartil son los 28 años, la mediana está en los 34 años y tercer cuartil está en los 41 años. Hay 272 observaciones con valores outliers para la variable edad. Los valores outliers para edad empiezan a partir de los 61 años. La moda para los valores outliers son los 61 años de edad.



Se puede observar la distribución de edades de los clientes varía un poco entre los clientes que no cayeron en default (valor 0) y los clientes que sí cayeron en default del pago de su tarjeta de crédito (valor 1). El primer cuartil, la mediana y el tercer cuartil para los clientes

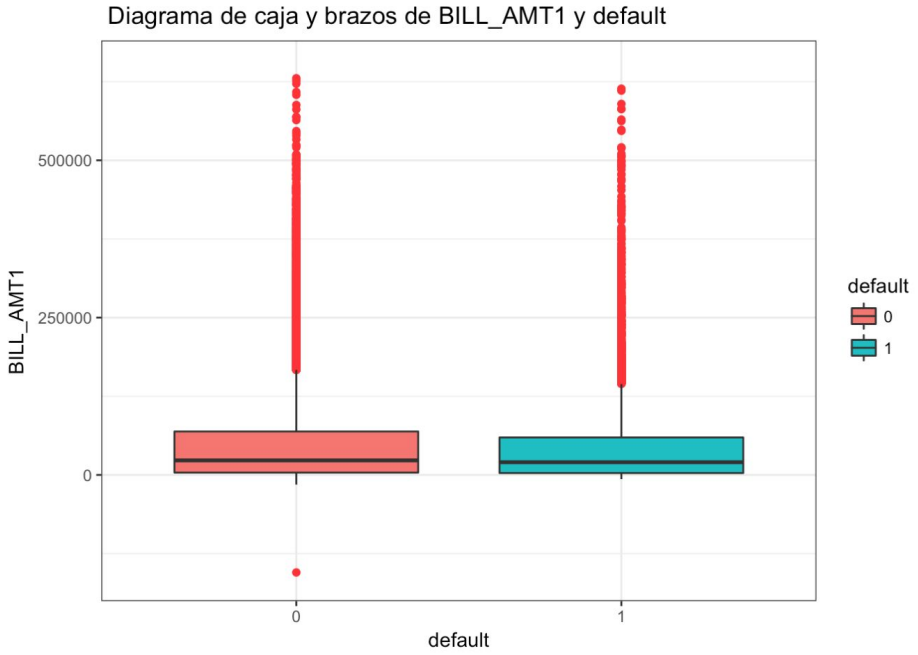
que no cayeron en default (valor 0, de color rosa) es menor que para los clientes que sí cayeron en default. Se podría decir que los clientes que no cayeron en default (valor 0) son un poco más jóvenes que los clientes que sí cayeron en default. Además, los valores outliers para los clientes que no cayeron en default empiezan en el 61, mientras que los outliers para los clientes que sí cayeron en default empiezan en valores mayores a los 61 años.

Variable Bill_Amount1



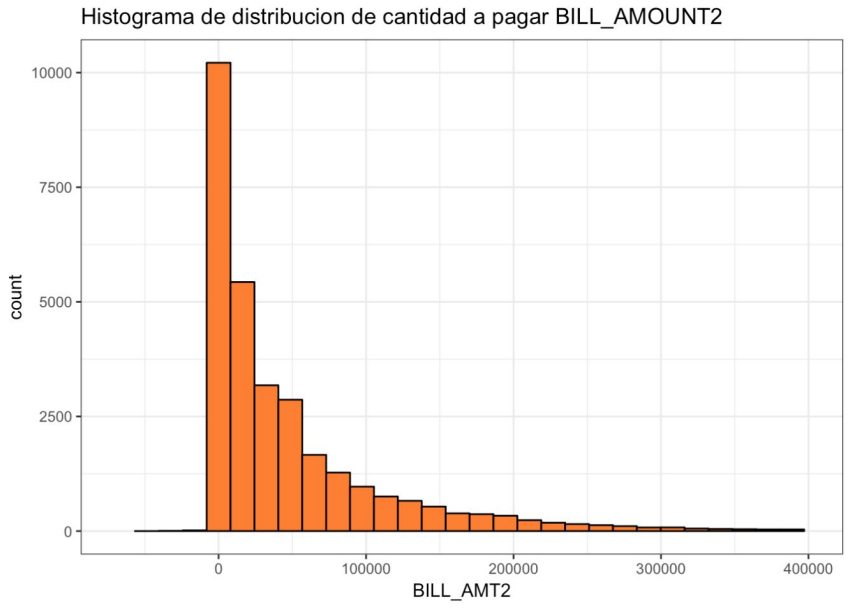
El valor máximo de Bill_Amount1 es 964,511, por lo que se decidió hacer "zoom" en la gráfica para poder analizarla de una manera más clara. Se puede observar que la distribución para el monto a pagar para el mes de septiembre de 2005 es parecida a la distribución exponencial. La mayoría de los valores se concentran entre el 0 y 100,000: esto hace sentido considerando que el tercer cuartil se marca por el valor 67,000.

Como se mencionó anteriormente, la distribución se concentra entra 3558 (primer cuartil) y 67,087 (tercer cuartil). Los valores outliers son valores negativos (saldos a favor) y valores positivos. Hay 2400 observaciones con valores outliers, de los cuales 2370 son valores únicos.

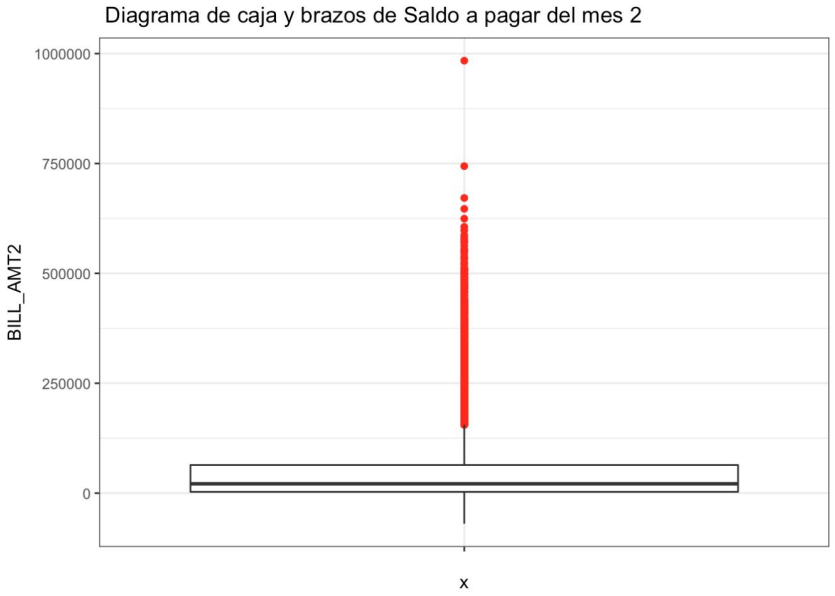


La distribución de Bill_amount1 para los clientes que no cayeron en default contaba con un valor outlier cercano a 1’000,000 y a 750,000. Para ver con mayor claridad la gráfica, hicimos zoom. Se puede observar que la distribución de los clientes que no cayeron en default (color rosa) está un poco más dispersa. Además los clientes que no cayeron en default son los que cuentan con saldos a favor (valores menores a cero).

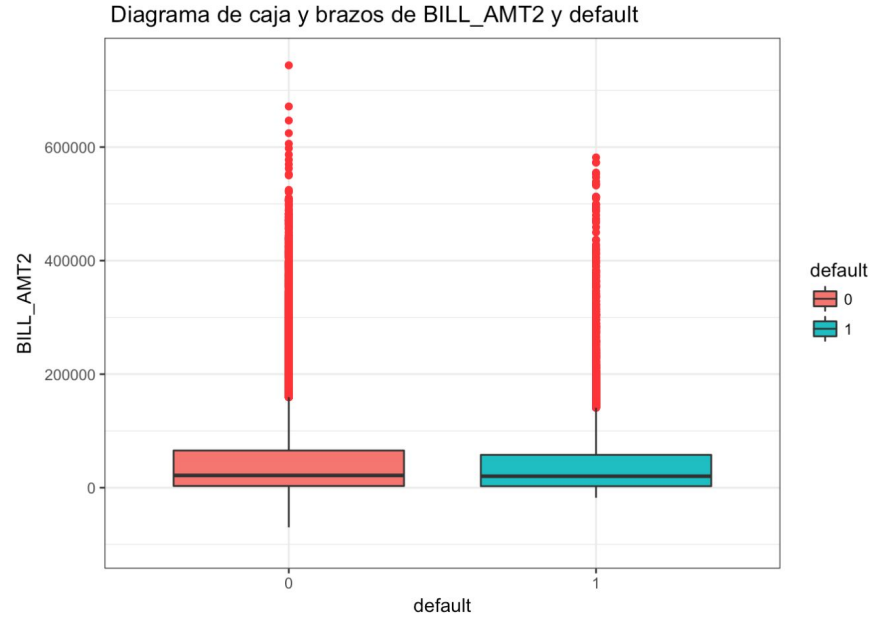
Variable Bill_Amount 2



Se puede observar que la distribución para esta variable también es parecida a una distribución exponencial. La mayoría de los datos están concentrados entre 0 y 100,000. Esto hace sentido pues el tercer cuartil de esta variable es 64000 aproximadamente.

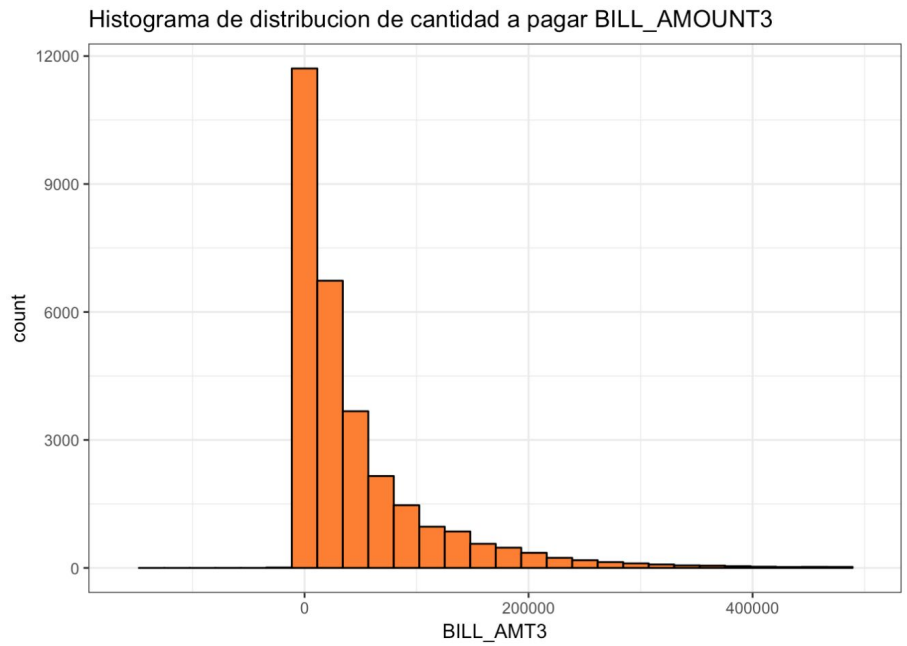


Como se mencionó anteriormente la distribución de esta variable está centrada entre 2984 y 64,000, que son primer cuartil y tercer cuartil respectivamente. Se puede observar que hay muchos valores outliers. Hay 2394 observaciones de esta variable que son valores outliers. Los valores outliers empiezan a partir 155,635.

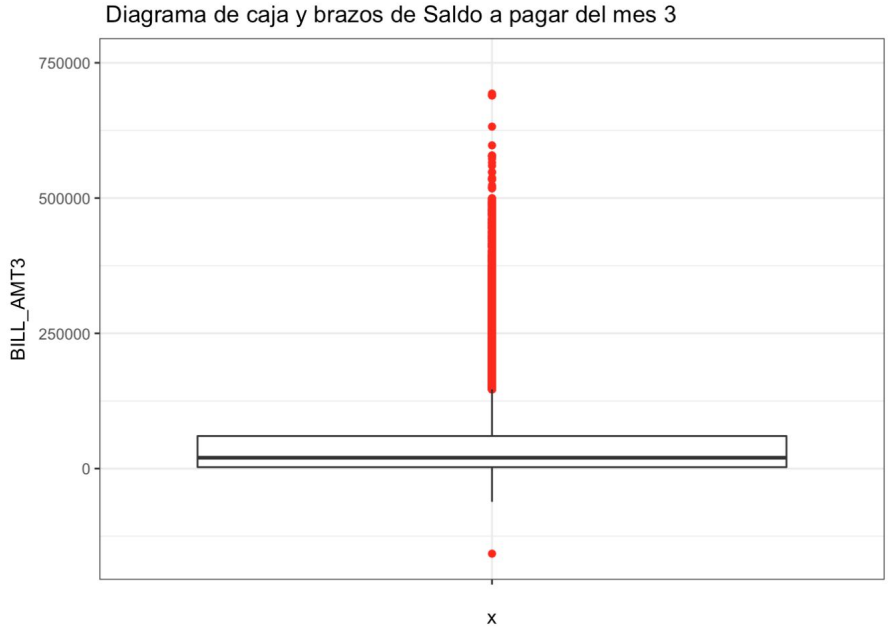


Esta variable también contaba con un valor outlier cerca de 1’000,000 y de 750,000 por lo que decidimos hacer zoom a la gráfica. La distribución de la cantidad a pagar para el mes 2 es muy parecida entre los clientes que cayeron en default (color azul) y los que no cayeron. Sin embargo,los clientes que no cayeron en default tienen valores outlier de mayor magnitud.

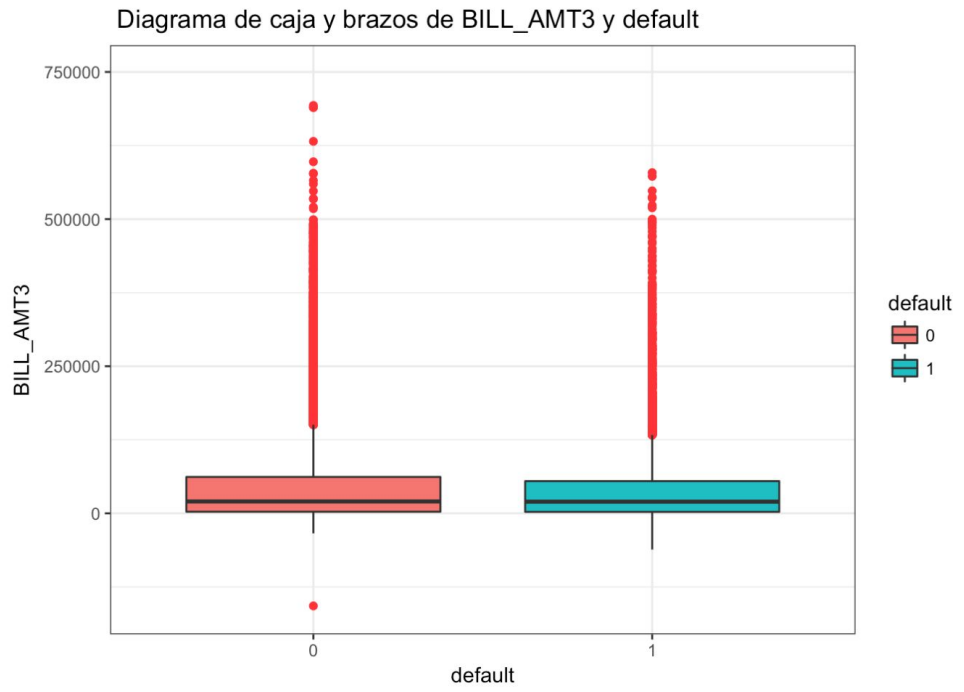
Variable Bill_Amount3



Nuevamente vemos una distribución parecida a la exponencial. Los valores se concentran entre 0 y 100,000.

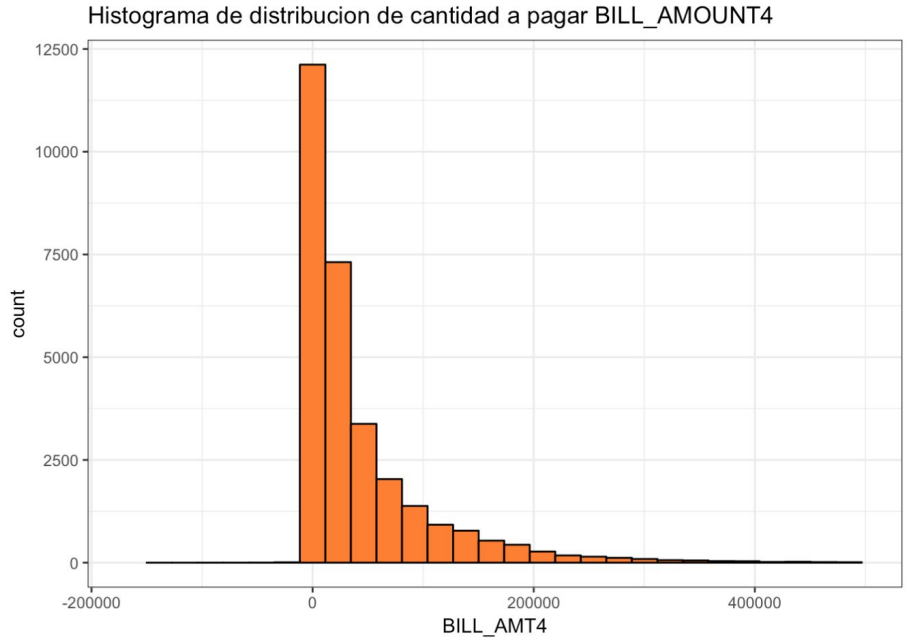


Se hizo "zoom" a la gráfica pues había valores outliers de gran magnitud (mayores a 1'500,000). La distribución se concentra entre 2666.25 y 60164.75. Hay valores outliers negativos y valores outliers positivos. Para esta variable, hay 2469 observaciones con valores outliers. El mínimo de los valores outliers es -157264, que también es el mínimo global.

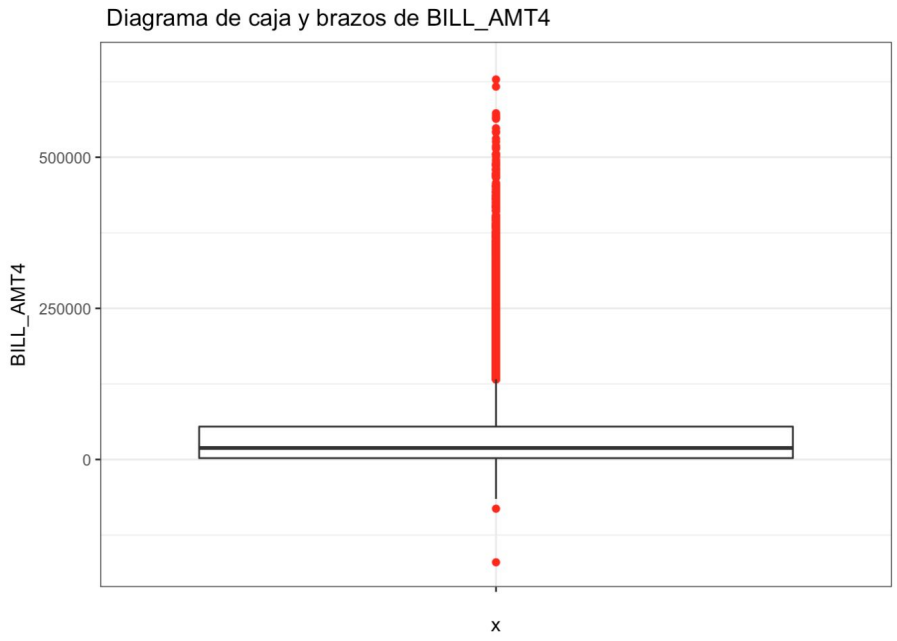


Vemos que las distribuciones para los clientes que cayeron en default (color azul) y para los que no cayeron (color rosa) es bastante similar. La principal diferencia es que algunos valores outliers para los clientes que no cayeron en default (valor 0, color rosa) son de mayor magnitud.

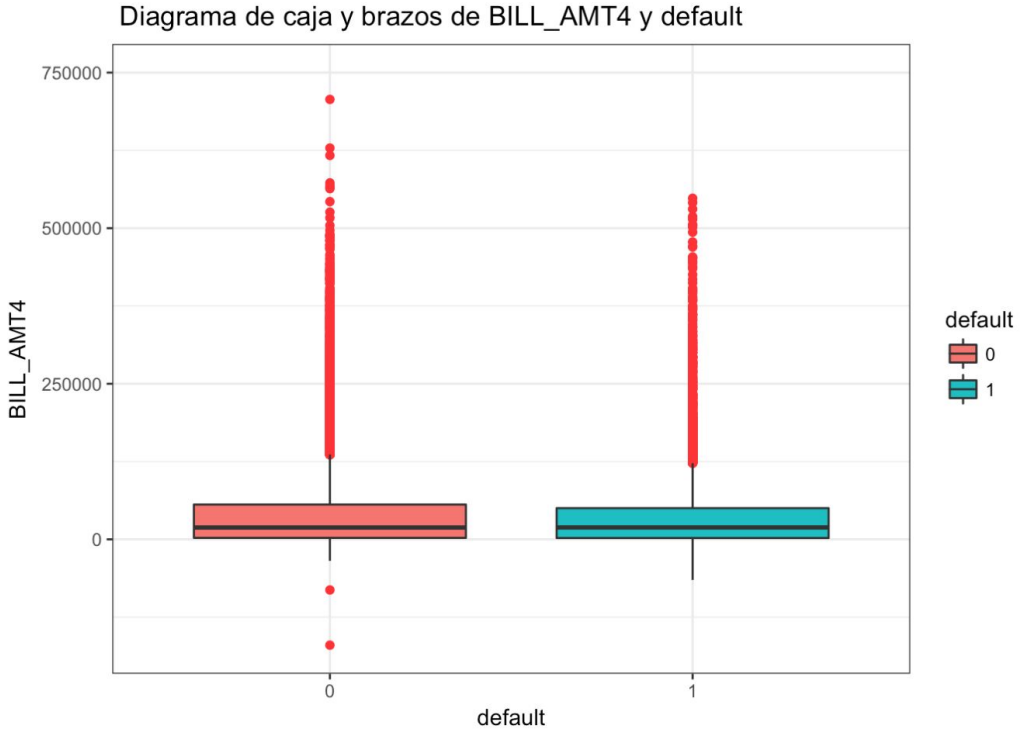
Variable Bill_Amount4



Nuevamente, se puede observar una distribución parecida a una distribución exponencial. La mayoría de los valores se concentran entre 0 y 100,000.

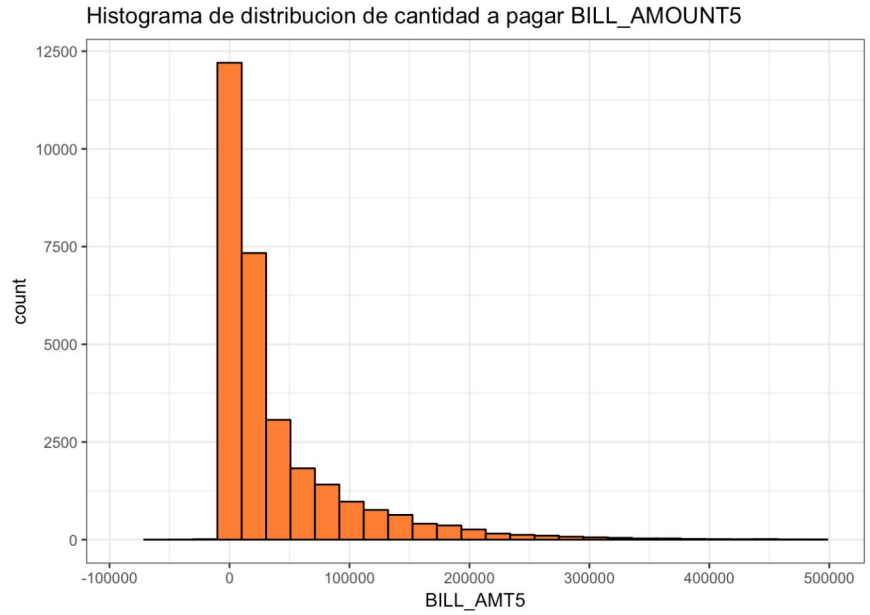


Se puede observar que la distribución se concentra entre 2323.00 (primer cuartil) y 54470.00 (tercer cuartil). Los valores outliers empiezan en 125,000 aproximadamente. Hay valores outliers positivos y negativos. Hay 2623 observaciones con valores outliers, el valor mínimo es 170000 (que también es el mínimo global).

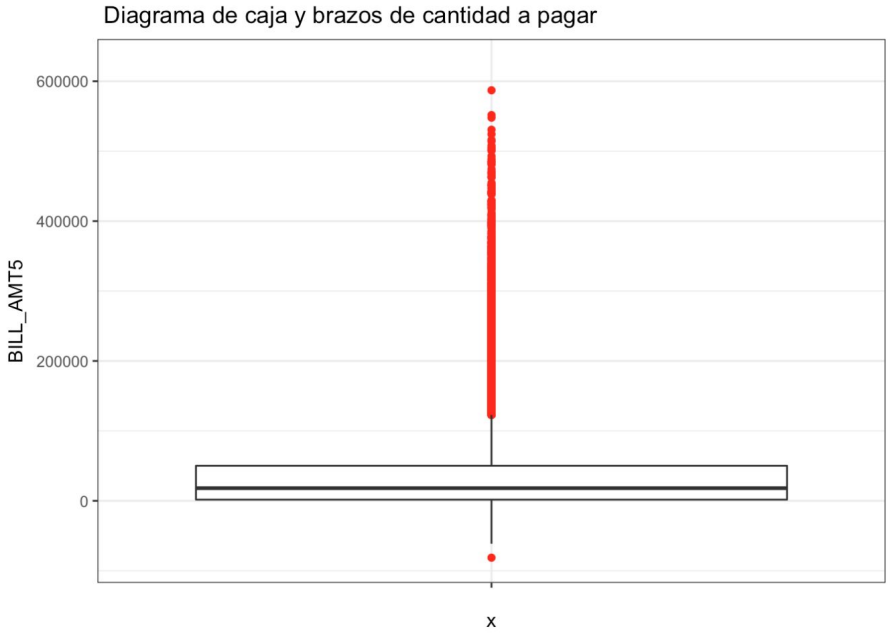


Se puede observar que la distribución para el monto a pagar en el mes 4 es bastante similar para los clientes que no cayeron en default y para los que sí cayeron. Sin embargo, los clientes que no cayeron en default tienen valores outliers negativos y de mayor magnitud.

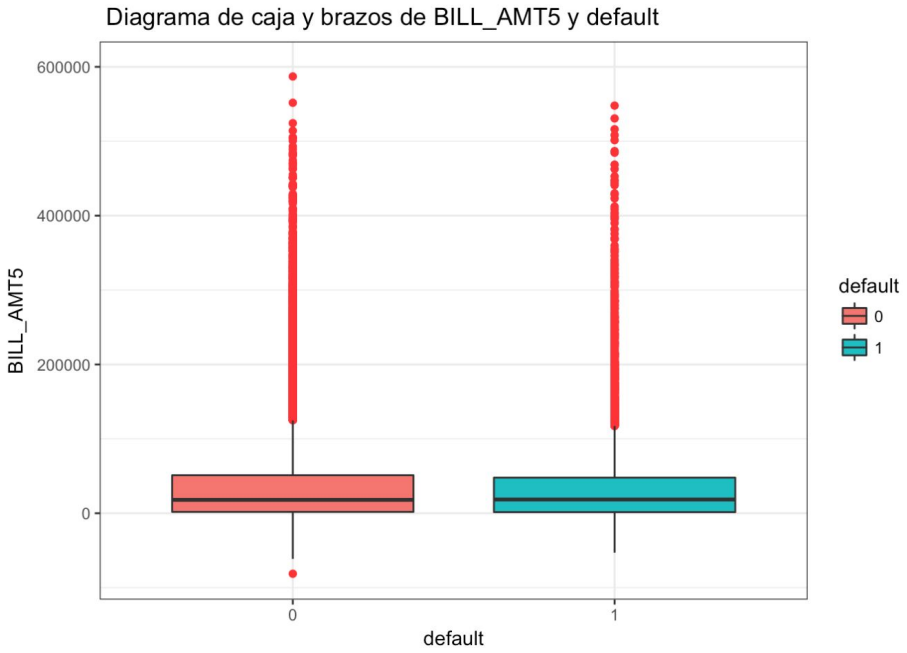
Variable numérica Bill_Amount5



Nuevamente, se puede observar una distribución parecida a la distribución exponencial. Gráficamente, la mayoría de los datos se concentran entre 0 y 100,000.

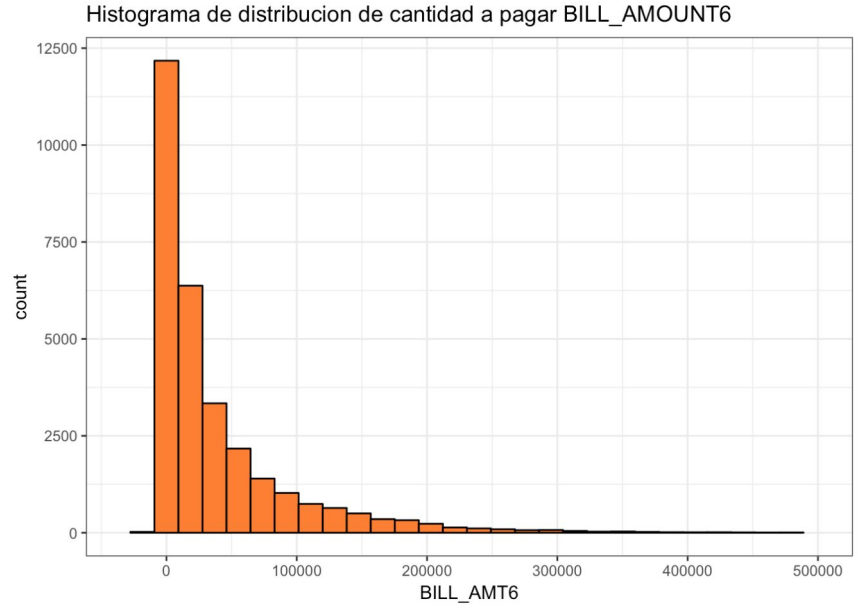


La distribución se concentra entre 1761 (primer cuartil) y 50149.75 (tercer cuartil). La mediana es 18097.0. Hay valores outliers negativos, y los valores outliers no negativos empiezan a partir del 100,000 aproximadamente. Hay 2724 observaciones con valores outliers para esta variable. Para los outliers negativos el valor mínimo es -81334, que también es el mínimo global.

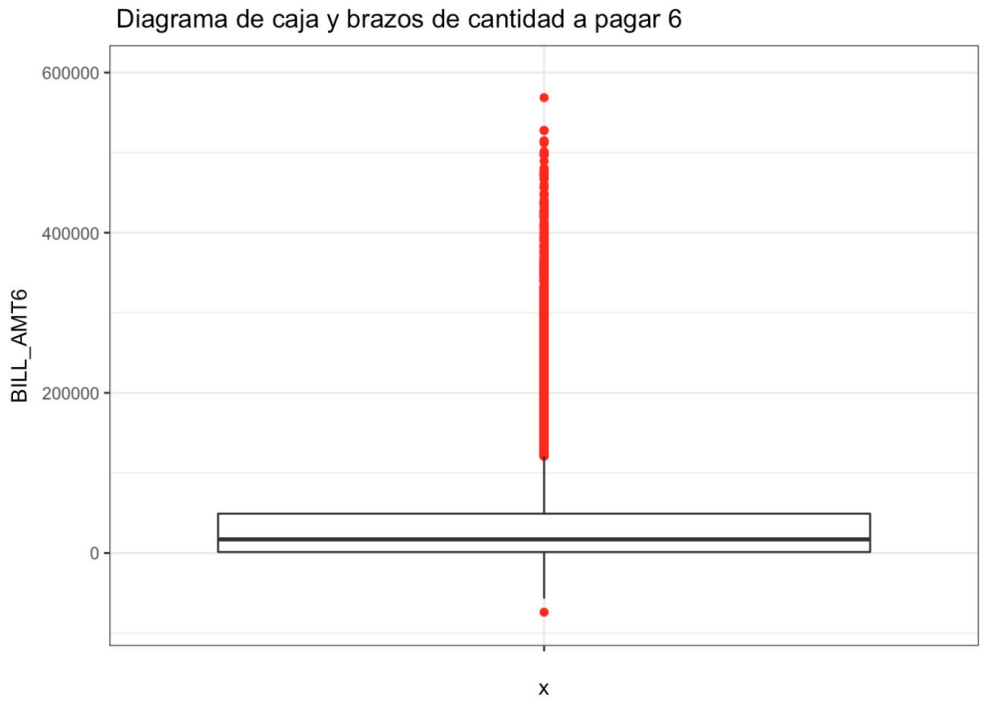


Se puede observar que la distribución del pago para el mes 5 es bastante similar para clientes que cayeron en default y para los que no cayeron.

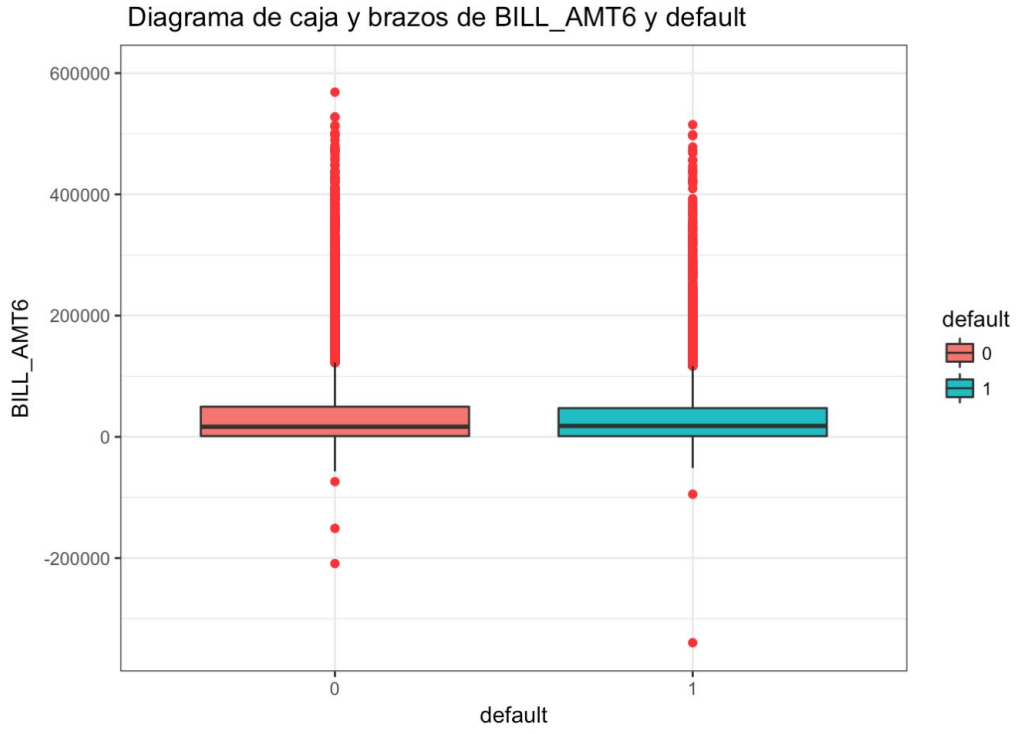
Variable Bill_Amount6



Nuevamente se puede observar una distribución parecida a una distribución exponencial. La mayoría de datos se concentran entre 0 y 50000.

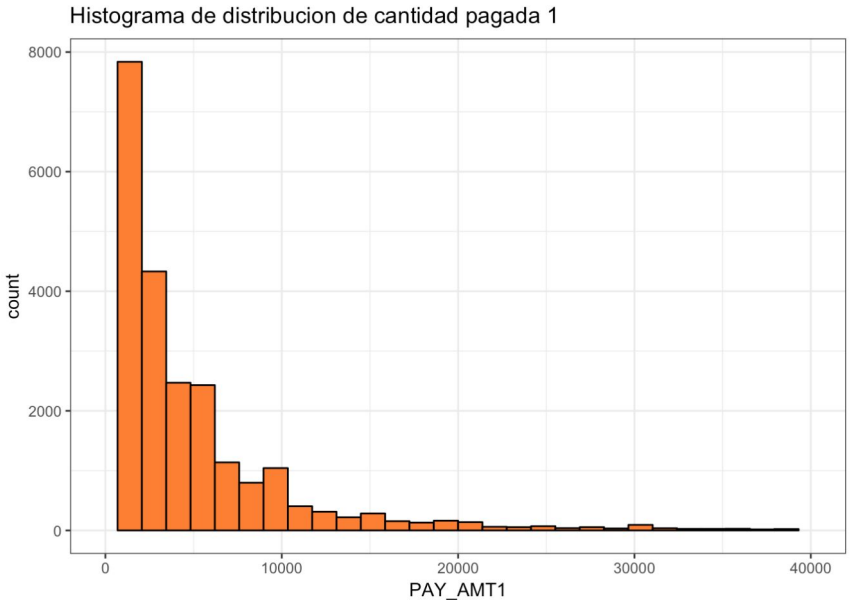


Se puede observar que la distribución se concentra entre 1256 (primer cuartil) y 49146.50 (tercer cuartil). La mediana es 17067.0. Hay 2696 observaciones con valores outliers para esta variable. El valor negativo de mayor magnitud es 339603, que también es el mínimo global.

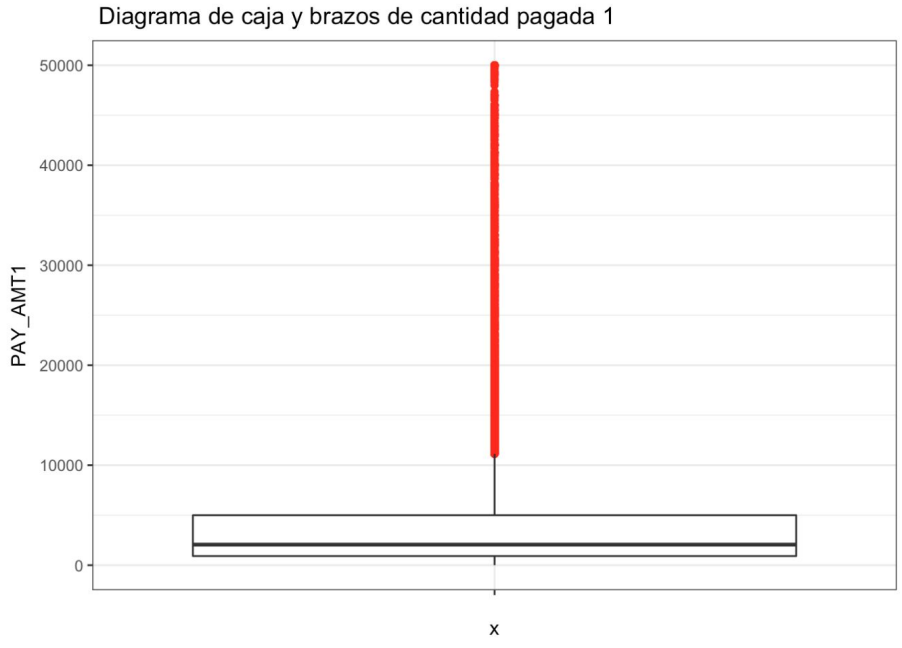


Se puede observar que la distribución para la cantidad a pagar del mes 6 para los clientes que cayeron en default y para los que no cayeron es bastante parecida.

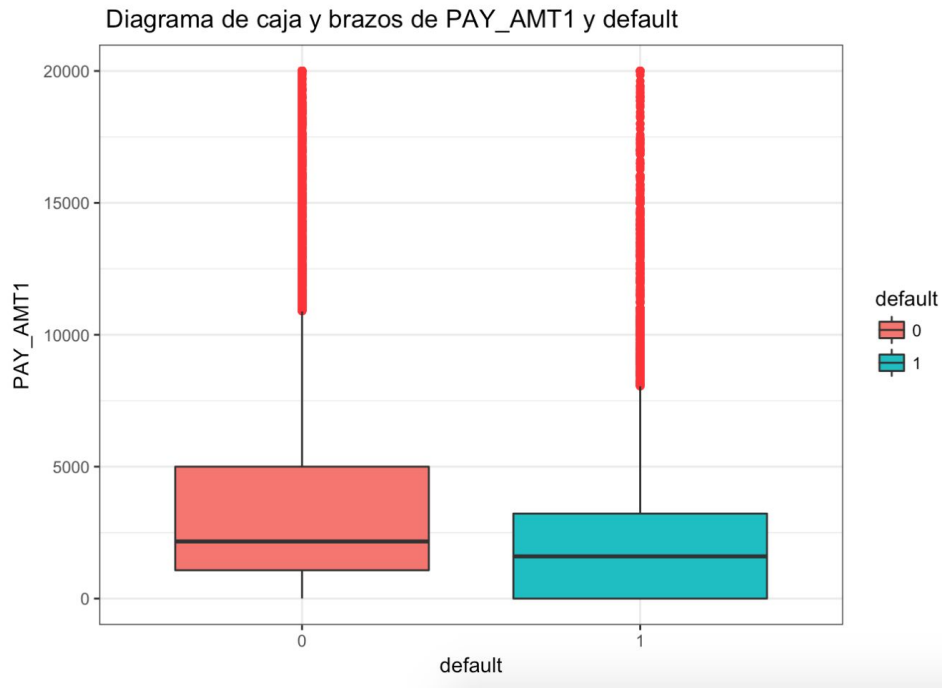
Variable Pay_Amount1



Se puede ver que en la distribución para los pagos del mes 1, la mayoría de las observaciones se concentran entre 0 y 5000.

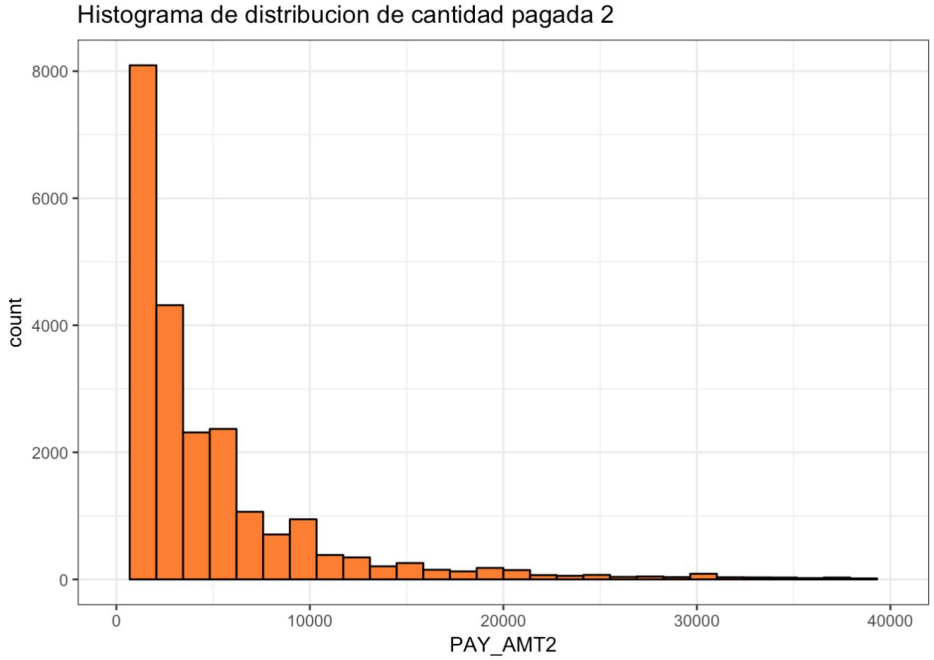


Se puede observar que la distribución de los pagos para el mes 1 se concentran entre 1000 y 5000 (primer y tercer cuartil respectivamente). La mediana es 2100. Los valores outliers empiezan a partir del valor 10,000 aproximadamente. Para esta variable hay 2729 observaciones con valores outliers. La magnitud del outlier más chico es de 11,016.

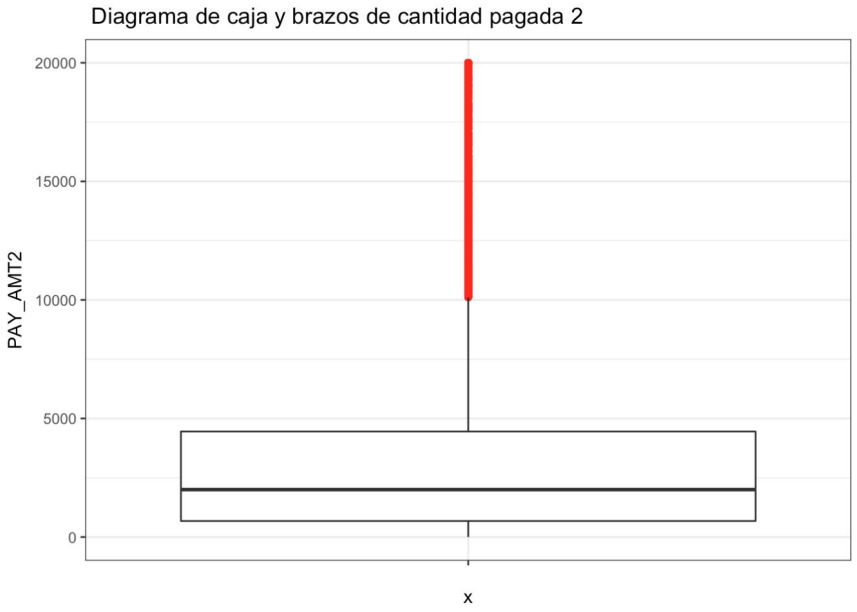


Hay valores outliers de gran magnitud, pero se decidió hacer "zoom" a la gráfica para poder analizar de manera más detallada cómo difieren las distribuciones de pagos para clientes que cayeron en default y para los que no cayeron. Se puede ver que los pagos de los clientes que sí cayeron en default son menores a los clientes que no cayeron en default.

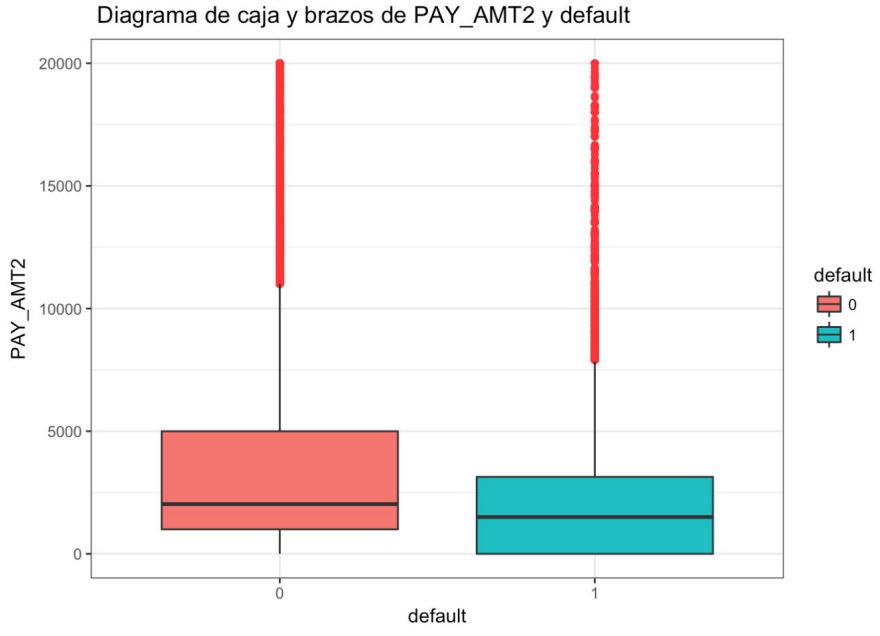
Variable Pay_Amt2



Nuevamente, se necesitó hacer "zoom" a la gráfica pues el valor máximo es 1'684,259.Se puede observar que la mayoría de los datos se concentran entre 0 y 5000.

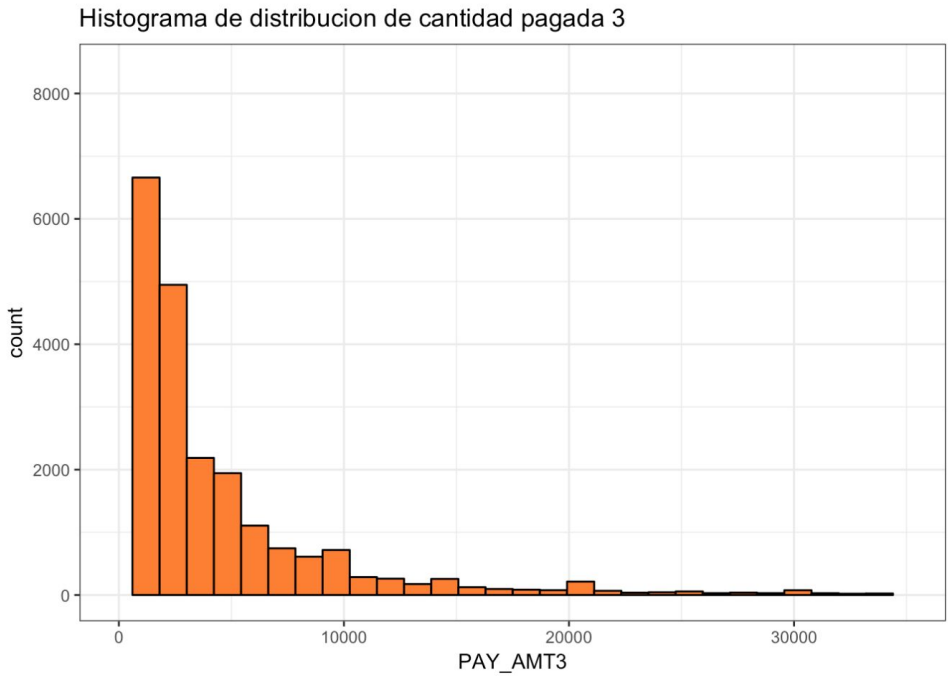


El outlier de mayor magnitud para esta distribución era mayor a 1’500,000. Nuevamente, fue necesario hacer “zoom” a la gráfica para apreciar de manera adecuada la distribución. Se puede observar que la distribución se concentra entre 0 y 5000. El primer cuartil es 832 y el tercer cuartil 5000, mientras que la mediana es 2008. Se puede observar que los valores outliers empiezan a partir del 10000 aproximadamente. Hay 2969 observaciones que tienen valores outliers para esta variable. Los valores outliers empiezan a partir del valor 11253.

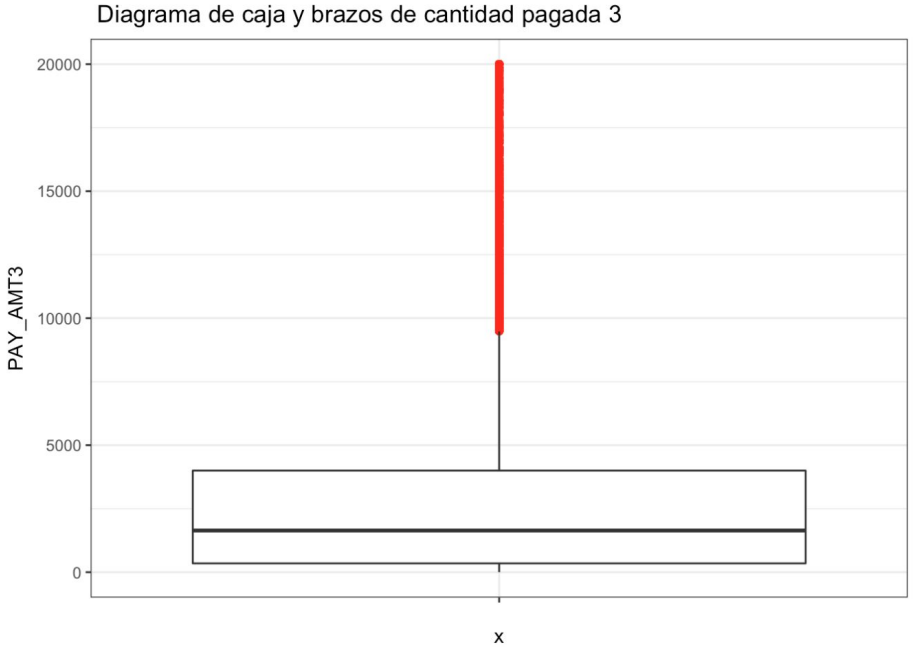


Vemos que los pagos de los clientes que sí cayeron en default son menores que los pagos de los clientes que no cayeron en default.

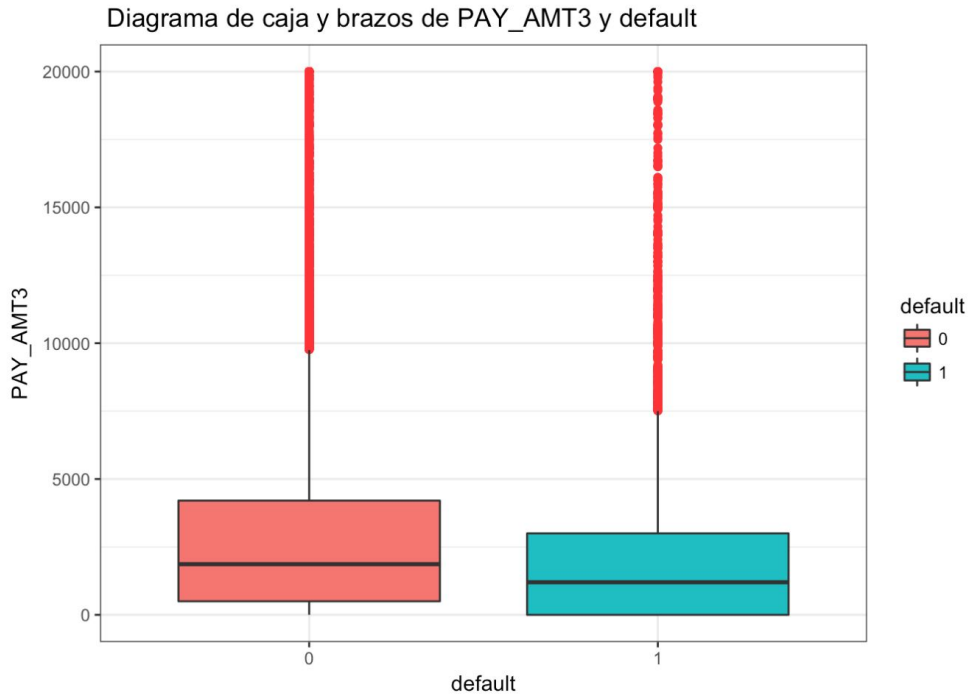
Variable Pay_Amt3



El valor máximo de pagos para este mes fue 896,040 lo cual provocaba que el rango de la gráfica fuera muy grande y no se pudiera apreciar bien la distribución. Se puede observar que la mayoría de las observaciones están concentradas entre 0 y 5000.

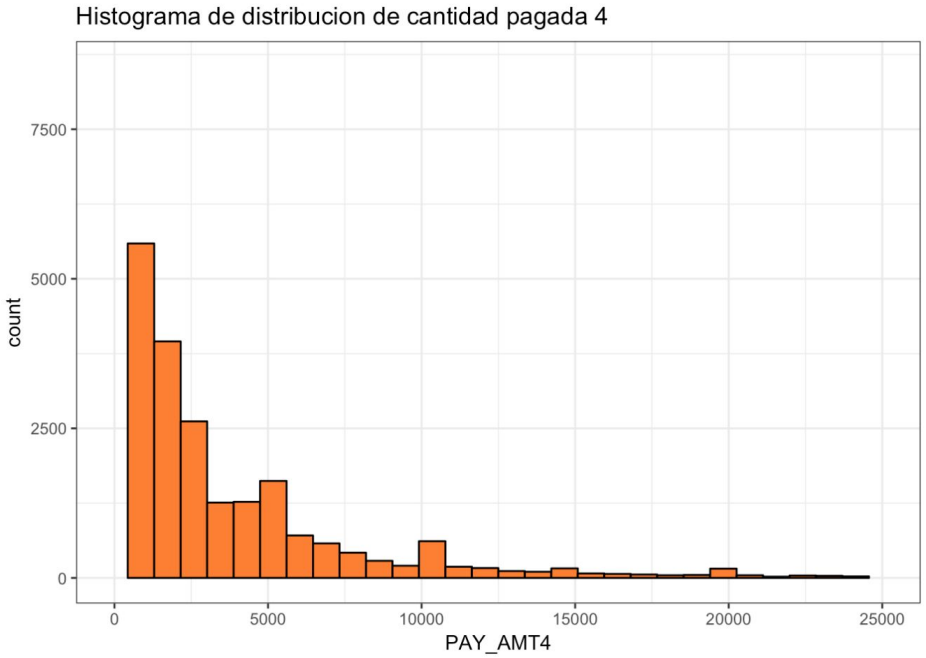


Hay valores outliers mayores a 750,000, por lo que nuevamente fue necesario hacer "zoom" a la gráfica. Se puede observar que la distribución se concentra entre el 0 y el 5000. El primer cuartil es 390, la mediana es 1800 y el tercer cuartil es 4500. Los valores outliers empiezan en valores cercanos a 10000. Hay 2591 observaciones con valores outliers para esta variable. Los valores outliers empiezan a partir del valor 10668.

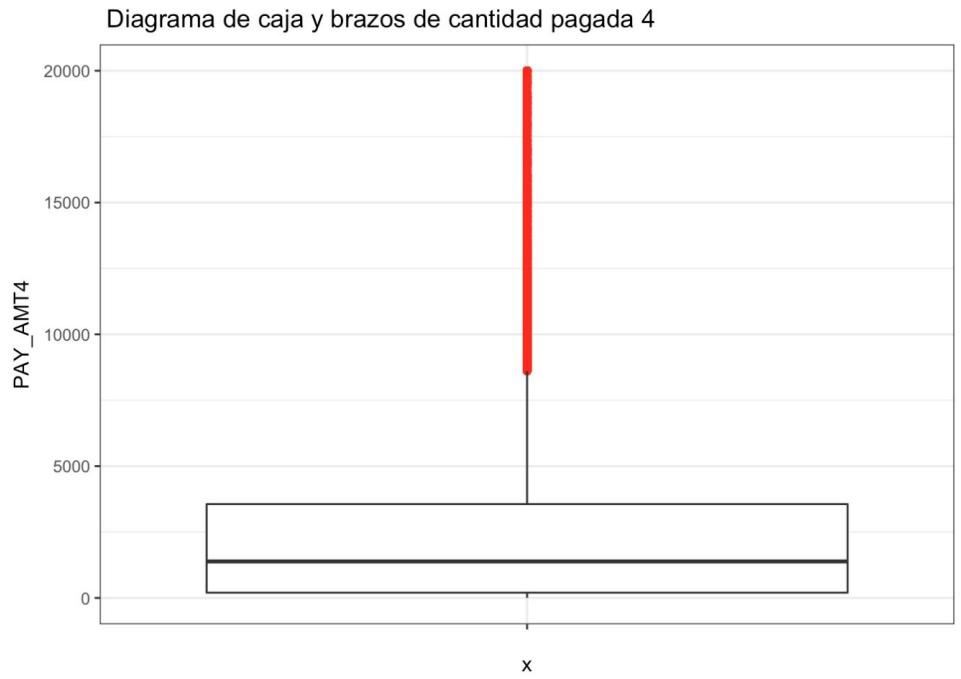


Nuevamente, se puede observar que los pagos de los clientes que sí cayeron en default (color azul) son menores que los pagos de los clientes que no cayeron en default (color rosa).

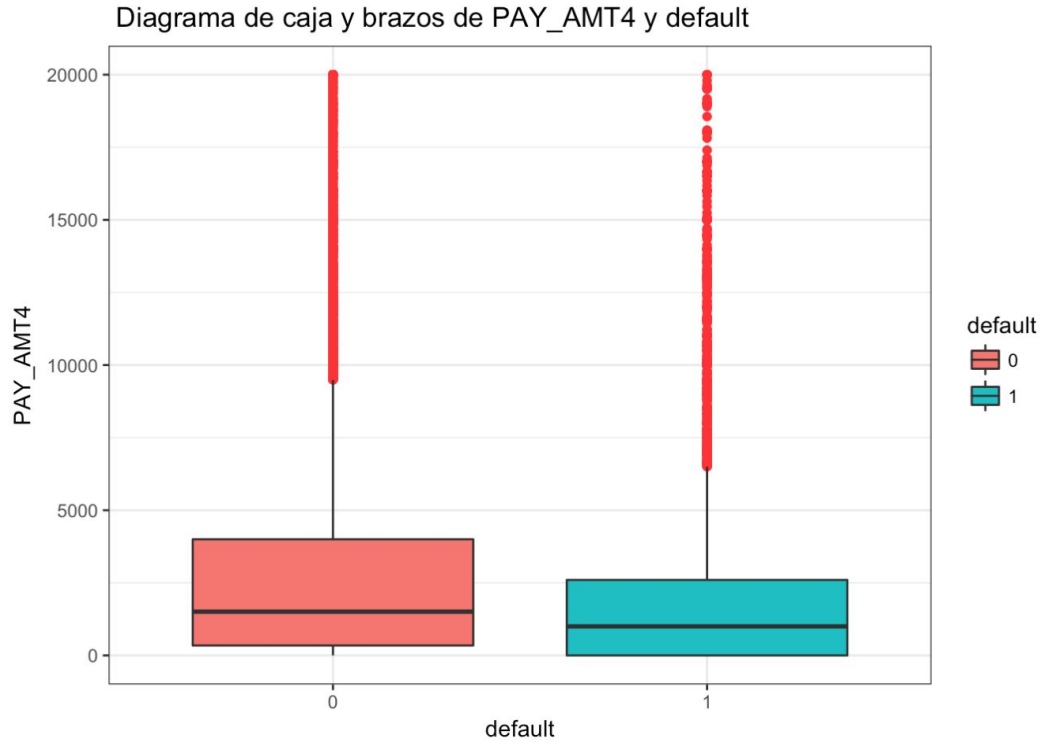
Variable numérica Pay_Amt4



El valor máximo para la variable es 621,000 por lo que fue necesario hacer zoom a la gráfica para apreciar mejor su distribución. Se puede observar que la mayoría de las observaciones se concentran entre 0 y 5000.

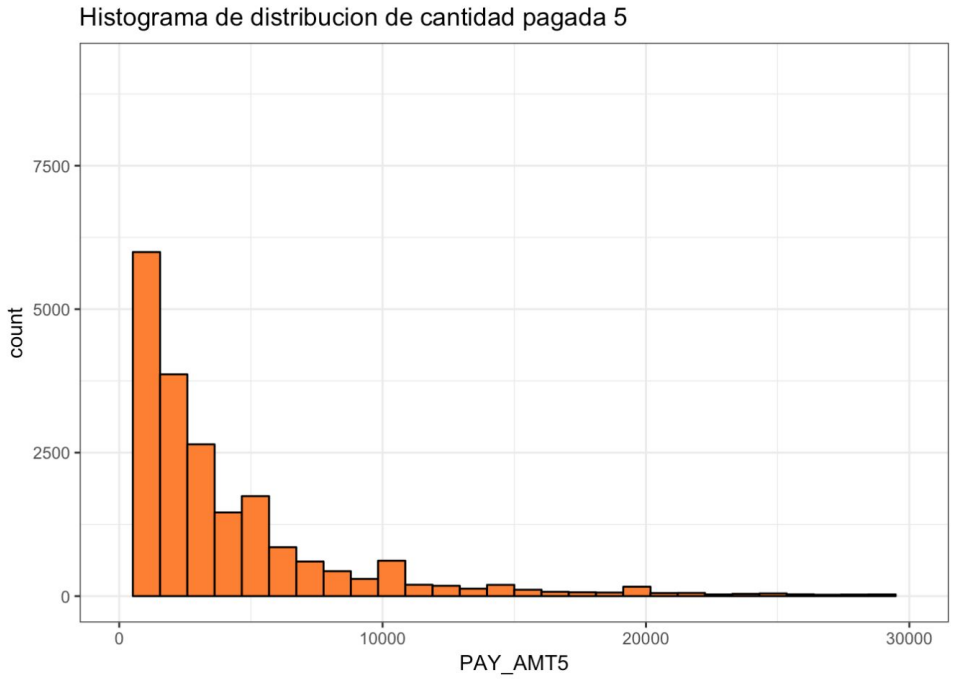


La distribución se concentra entre 296 (primer cuartil) y 4013 (tercer cuartil). La mediana es 1500. Hay 2994 observaciones con valores outlier para esta variable y las observaciones outliers empiezan a partir del valor 9590 aproximadamente.

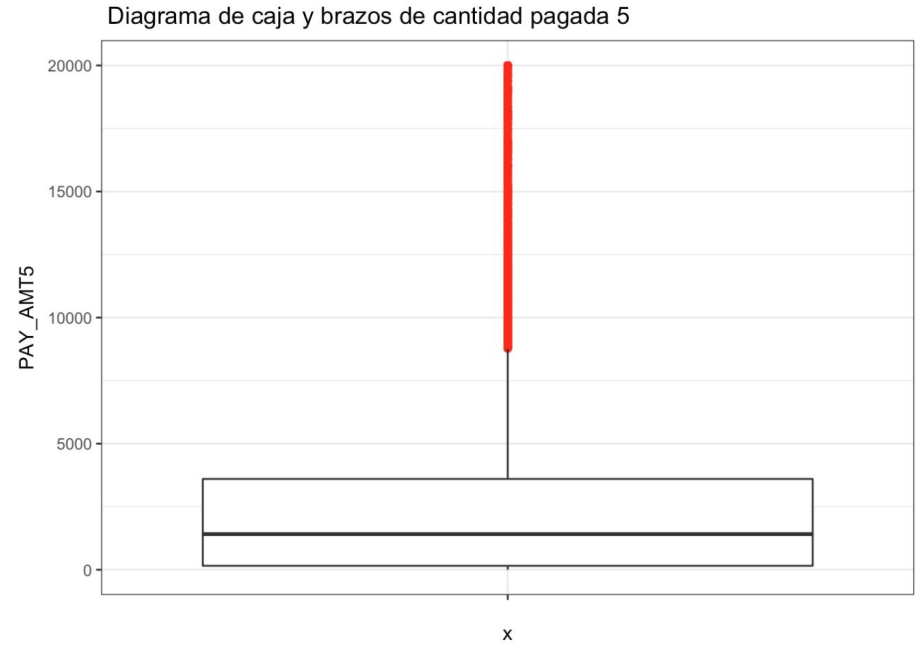


Nuevamente, se puede ver que la distribución y cuartiles de los pagos de clientes que sí cayeron en default son menores que los clientes que no cayeron.

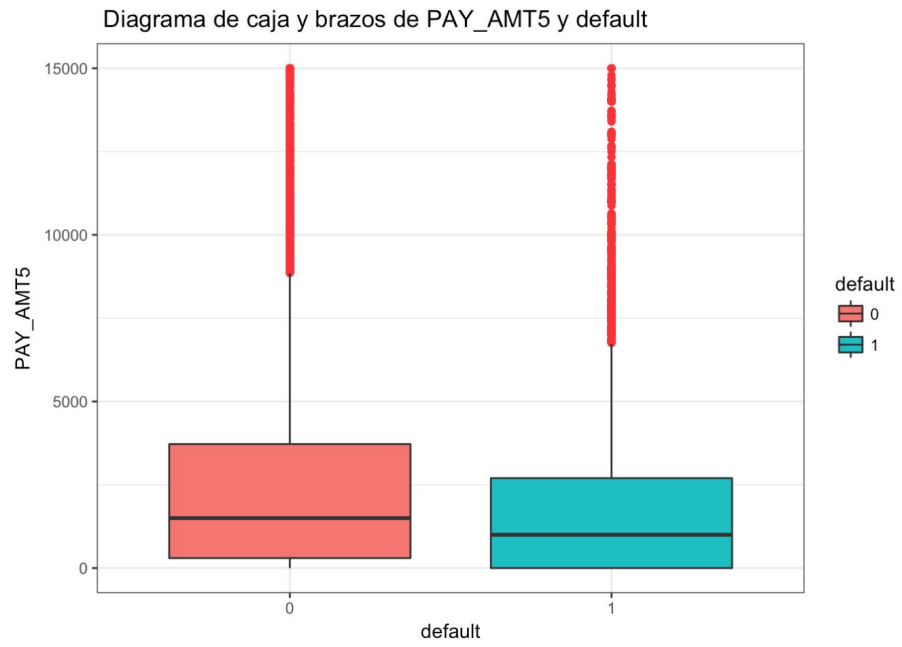
Variable Numérica Pay_Amount5



Se puede observar que la mayoría de los datos se concentran entre 0 y 5000. El valor máximo es de 426,529, por lo que fue necesario hacerle zoom a la gráfica.

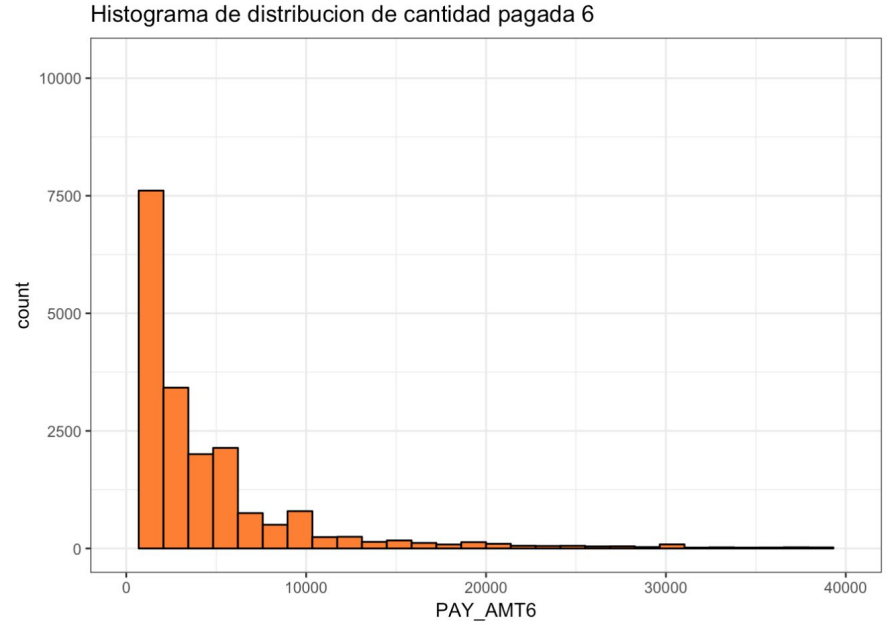


Se puede observar que la distribución para esta variable se concentra entre el 252 (primer cuartil, valor cercano al cero dado el rango de la variable), y el valor 4031 (tercer cuartil). La mediana para esta variable es 1500. Hay 2944 observaciones con valores outliers para esta variable, y los valores outliers empiezan a partir del valor 9718.

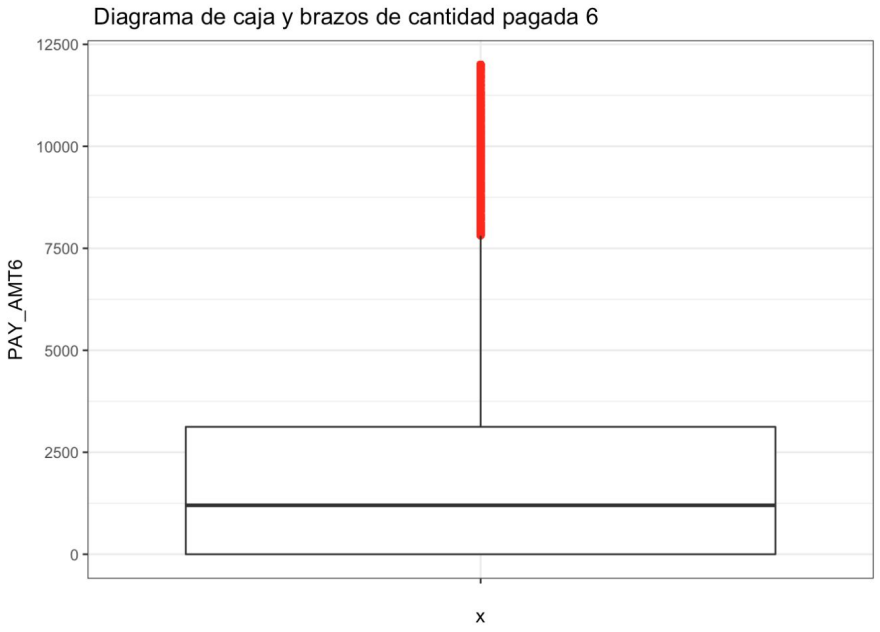


Nuevamente, se puede observar que la distribución de pagos para los clientes que sí cayeron en default es menor que para los clientes que no cayeron.

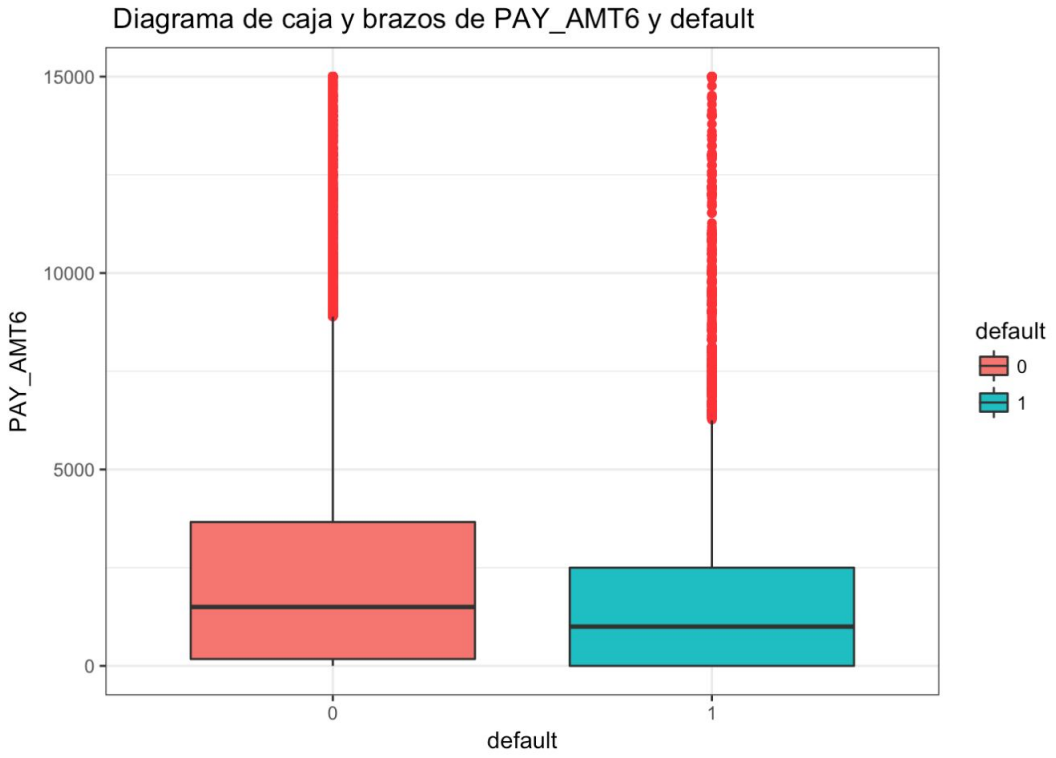
Variable Numérica Pay_Amount 6



El valor máximo para esta variable es 528,666 lo cual hacía que todas las observaciones estuvieran centradas en cero dada la escala original de la gráfica. Se puede observar que la mayoría de las observaciones están concentradas entre 0 y 5000.



Por la gran dispersión de los datos, fue necesario hacer zoom a la gráfica nuevamente. Se puede observar que la distribución se concentra entre 116 (primer cuartil) y 4000 (tercer cuartil). La mediana de los datos es 1500. Hay 2946 observaciones con valores outliers para esta variable. Los valores outliers empiezan a partir del valor 9832.



Al igual que en las otras variables, se observa que la distribución de pagos para los clientes que sí cayeron en default es menor que para los clientes que no cayeron en default.

Observaciones generales de las variables numéricas

En las variables numéricas del tipo Bill_AmtNum y Pay_AmtNum se pudo observar un rango (diferencia entre valor máximo y mínimo) muy grandes. En ambos tipos de variables hay más de 2,000 observaciones con valores outliers. Dado que la variable default es la variable target, se analizó la relación de esta variable con las variables continuas. Se observó lo siguiente:

Para la distribución de la variable Limit_Bal (límite de crédito) sí hay diferencia en distribución para clientes que cayeron en default y para los que no cayeron. En general, el límite de crédito de los clientes que sí cayeron en default es menor que para los clientes que no cayeron.

Para la variable Age, no hay mucha diferencia entre las distribuciones de los clientes que sí cayeron en default y los que no cayeron. El primer cuartil, mediana y tercer cuartil de los clientes que sí cayeron en default es un poco mayor que para los clientes que no cayeron.

Para las variables de tipo Bill_AmtNum (cantidad a pagar) realmente no había mucha diferencia entre los clientes que cayeron en default y los que no cayeron.

Para las variables de tipo Pay_AmtNum (cantidad pagada) sí se notaba una diferencia entre los clientes que sí cayeron y los que no cayeron. En general, los clientes que sí cayeron en default (valor 1 para la variable) hicieron pagos por montos menores.

Aplicación de modelos de clasificación

Se van a aplicar distintos modelos de clasificación para poder predecir si un cliente podrá pagar su tarjeta de crédito o no dadas ciertas variables categóricas y continuas. Como se observó en el análisis exploratorio, no todas la variables varían según el valor de la variable default, por lo que se sospecha que hay variables que se pueden descartar porque no aportan información. A pesar de que en el análisis exploratorio se identificó que en las variables numéricas continuas hay más de 2000 observaciones con valores outliers, éstas no se van a eliminar. Las observaciones outliers se deben eliminar si hubo un error de medición, y no sabemos con certeza si esto pasó o no.

El primer paso fue eliminar los registros que tuvieran valores vacíos, al hacerlo el set tenía 29,927 observaciones; es decir, 97.76% del set original. Posteriormente, el siguiente paso fue separar el set de datos en entrenamiento, validación y pruebas. Se obtuvo una muestra aleatoria de tamaño 70% del set de datos para obtener el set de entrenamiento, el 30% de las observaciones restantes fue destinado al set de pruebas, y el 5% del set de entrenamiento fue usado como set de validación. El set de pruebas tenía 19,902 observaciones, el set de validación tenía 1,047 observaciones y el set de pruebas tenía 8,978 observaciones. Ya que los sets fueron definidos, cada uno se guardado en un archivo csv (comma-separated values) por separado: de esta manera se asegura que al momento de entrenar los modelos, éstos no se "contaminen" por información futura. El set de pruebas no fue usado hasta saber cuáles eran los parámetros “seleccionados” del modelo.

El primer modelo que se aplicó fue un árbol de decisión, la ventaja de este algoritmo es que nos dice cuánto aporta cada variable a la clasificación, por lo que es posible descubrir qué variables sí aportan información y cuáles no.

Posteriormente, se usó un modelo de ensamble random forest, que combina un número determinado de árboles de decisión. La ventaja del modelo random es que crea árboles diferentes, con diferentes parámetros, lo que fomenta la diversidad y aleatoriedad. Para hacer la selección final, se hace un voto de expertos. Por la naturaleza del algoritmo, éste modelo nos permite hacer selección de variables. Se aplicó el modelo con todas las variables y después con las variables más importantes y se compararon resultados para ver si convenía o no eliminar variables.

Después, se aplicó un Naive Bayes, por su simplicidad y simplemente para comparar resultados y el desempeño con los demás modelos. El último paso fue comparar los resultados con una red neuronal, por el poder que tiene este algoritmo para resolver problemas complejos.

Para cada modelo se justificarán los parámetros elegidos según aplique. Para cada modelo se usará la curva ROC (Receiving Operating Characteristic), su área bajo la curva y

una matriz de confusión para medir el desempeño. Los errores se clasificaron de la siguiente manera:

- True Positive : el modelo clasificó una observación como un cliente que sí cae en default y la clasificación real coincide.

- True Negative: el modelo clasificó una observación como un cliente que no cae en default y la clasificación real coincide.

- False positive: el modelo clasifica una observación como un cliente que sí cae en default y la observación real no coincide (el cliente realmente no cae en default).

- False negative: el modelo clasifica una observación como un cliente que no cae en default y la observación real no coincide (el cliente sí cae en default).

Como se mencionó anteriormente, el set de datos de obtuvo de Kaggle y fue provisto por UCI Machine Learning, por tanto no contamos con información real de los costos que trae consigo cada tipo de error. Si este fuera un análisis de datos para un cliente real, tendríamos que preguntar que tipo de error quisiera optimizar. La intuición nos dice que se debe escoger un punto de corte tal que se maximice el número de True Positives y se minimice el número de False Negatives (creemos que es más grave decir que un cliente no cae en default cuando en realidad sí cae).

Todos los algoritmos y procesamiento de datos se hicieron en el ambiente de R Studio.

Árbol de decisión

Para aplicar el modelo, no fue necesario escalar los datos pues el algoritmo lo hace por sí solo. Lo único que se requiere es establecer la variable target como tipo factor. El paquete que se usa es el C5.0 de Ross Quinlan. Los parámetros que el modelo necesita son trials y el número de minCases. El número de trials se refiere al número de veces que se "parte" el set de entrenamiento. El parámetro de minCases se refiere a cuántos elementos (observaciones) deben estar en un nodo para ser considerado hoja (es decir, para dejar de dividir). Se va a correr el algoritmo con distintos números de minCases para escoger el que logre el mejor desempeño. El número de trials se quedará fijo en 10.

Árbol con minCases=3

Al correr este árbol y probar el desempeño con el set de validación se obtuvo un área bajo la curva ROC de 0.7571439.

Árbol con minCases=5

Ya que aumentamos el número de minCases a 5, el área bajo al curva ROC aumentó a 0.770508.

Árbol con minCases=7

Como el desempeño aumentó al incrementar el número de minCases, se decidió aumentarlo aún más. Sin embargo, el área bajo la curva ROC bajó a 0.7571439, misma que se había obtenido para el árbol con 3 min cases.

Árbol con minCases=10

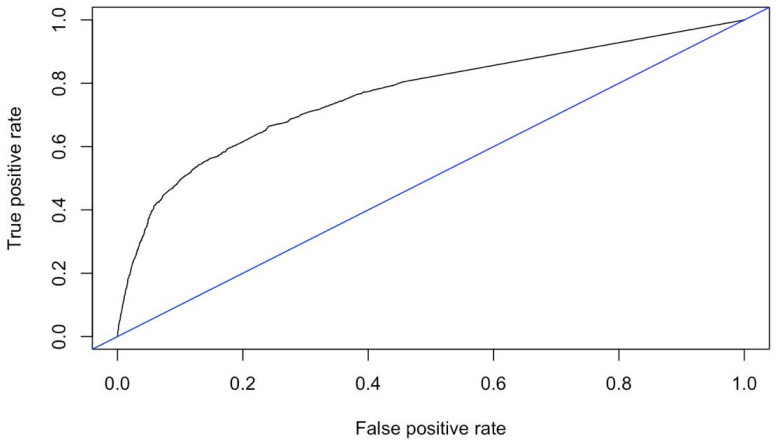
Por curiosidad, decidimos probar qué pasaba si se establecía minCases a 10. El área bajo la curva ROC nuevamente era de 0.7571439, la misma que para los árboles con 3 y 7 minCases.

Parámetros seleccionados

Como se detalló anteriormente, se corrieron 4 distintos árboles, cada uno con el mismo número de trials pero con un número de minCases distinto: 3,5,7,10. El número de minCases que obtuvo la mejor área bajo la curva fue minCases=5, por lo que ese es el valor seleccionado para ese parámetro.

Desempeño con el set de pruebas

El área bajo la curva ROC con el set de pruebas fue de 0.7644124, lo cuál representa una mejora con respecto a los resultados con el set de validación.



Curva ROC para el árbol de decisión con área bajo la curva de 0.765

Matriz de confusión

	Reference	
Prediction	**0**	**1**
0	6664	1330
1	306	678

En el set de pruebas, hay 6970 observaciones con clasificación real de 0 y 2008 observaciones con clasificación real de 1. Por tanto, se puede observar que obtuvo 678 observaciones True positives, 6664 observaciones True Negatives, 306 observaciones False Positives, 1330 de Falsos negativos. Se tienen las siguientes medidas de desempeño:

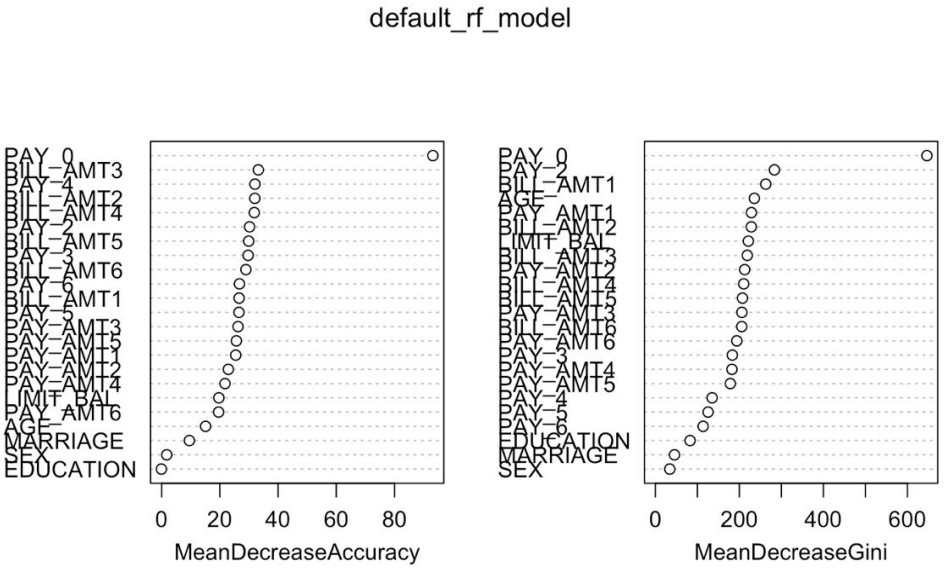
- Accuracy = 81.78%
- Precision = 68.90%
- Recall= 33.76%
- F- Measure= 0.4532086
- Punto de corte= 0.4863387

Como nuestro estudio busca maximizar el número de true positives y minimizar el número de false positives, las medidas de precision y recall son las que más importantes. El punto de corte seleccionado es 0.4863387, pues con este punto de corte tenemos una proporción de 0.3725 de True Positives, 0.04992 de Falsos positivos, 0.95 de True Negatives y 0.6275 de False Negatives. Se escogió este punto de corte pues minimiza los falsos positivos y maximiza los true negatives. En general, el modelo es mucho mejor clasificando las observaciones con default cero, lo cual hace sentido pues el set de datos está desbalanceado y casi el 75% de los datos cuenta con clasificación 0 y el resto con clasificación 1. Por tanto, si el resultado del modelo es mayor a 0.4863387, la observación se clasificará como default 1 (sí cae en default).

Random Forest

El primer paso para aplicar el modelo (sobre el set de entrenamiento) fue convertir todas las variables categóricas a tipo factor. Esto porque el hecho de que una categoría tenga un valor numérico mayor no implica que sea mayor. Por ejemplo, en la categoría SEX, los hombres tienen en valor 1 asignado y las mujeres tienen el valor 2 asignado, y esto no implica que una categoría valga más que otra. Se hizo una copia del set de entrenamiento, validación y prueba para poder convertir las variables categóricas a tipo de factor sin afectar los datos para los siguientes modelos. Los parámetros para este modelo son el número de árboles y el número de observaciones requeridas para que un nodo se convierta en hoja y el algoritmo se detenga y no siga dividiendo ese grupo de datos.

Modelo con todas las variables, 300 árboles y 10 observaciones como node size



Se puede observar que las variables que mayor ganancia de información (medidas por cómo se disminuiría el Gini de información si las eliminamos) al momento de intentar clasificar la variable default son Pay_0, Pay_2, Bill_AMT1, Age, Pay_Amt1, Bill_AMT2 , Limit_Bal , Bill_AMT3 , Pay_Amt2 , Bill_AMT4 , Bill_AMT5,Pay_Amt3, Bill_AMT6 y Pay_Amt6 (14 variables).

Desempeño con el set de validación

El área bajo la curva ROC para el set de validación de este random forest con todas las variables fue de 0.7773172, lo cual es un poco más alto que el área bajo la curva para el árbol de decisión con 5 minCases.

- Accuracy = 0.8156638
- Precision = 0.6223776
- Recall = 0.3903509
- F-Measure= 0.4797844

Modelo con selección de variables, 300 árboles y 10 observaciones como node size

Las variables que no aportan información (porque el MeanDecreaseGini es muy bajo) son SEX, MARRIAGE, EDUCATION, PAY_6,PAY_5, PAY_4,PAY_3, PAY_AMT5, PAY_AMT4. Por tanto, se corrió el modelo eliminando estas variables para analizar si el desempeño mejoraba o no.

Desempeño con el set de validación

El área bajo la curva al eliminar variables fue de 0.7560729, lo cual es menor al área bajo la curva del modelo que incluye todas las variables.

- Accuracy = 0.8147087
- Precision = 0.6214286
- Recall = 0.3815789
- F-Measure= 0.4728261

Se puede observar que las medidas de desempeño disminuyen un poco al eliminar variables. Por lo que se decide no quitar variables predictoras.

Modelo con todas las variables, 300 árboles y 5 observaciones como node size

Ya que se probó el desempeño con todas las variables y con selección de variables, se volvió a correr el modelo con todas las variables y 300 árboles pero cambiando el parámetro de node size a 5 observaciones.

Desempeño con el set de validación

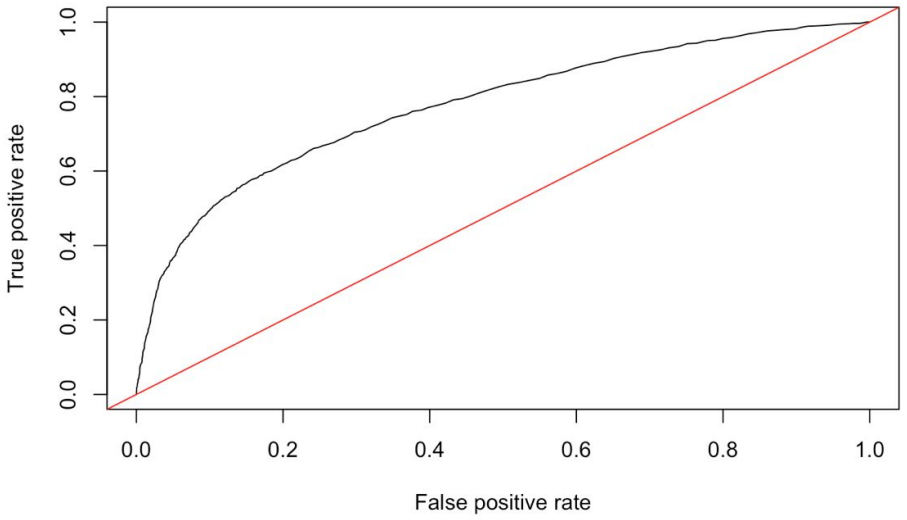
El área bajo la curva ROC fue de 0.777103, lo cual es un poco más bajo que para el modelo con 10 de node size. Las medidas de desempeño fueron las siguientes:

- Accuracy= 0.8204394
- Precision=0.6369863
- Recall= 0.4078947
- F-Measure= 0.4973262

Las medidas de desempeño presentaron una mejora con respecto al modelo con node size de 10. Por tanto, el modelo que usamos para probar el desempeño con el set de pruebas es el modelo random forest con 300 árboles y node size de 5.

Desempeño con el set de pruebas

El área bajo la curva fue de 0.7753478, lo cual es un poco menor que para el set de validación pero mayor que para el árbol de decisión.



Curva ROC para el el modelo random forest con área bajo la curva de 0.775

Matriz de confusión

	Reference	
Prediction	**0**	**1**
0	6574	1221
1	395	787

Las medidas de desempeño son las siguientes:

- Accuracy= 0.8199844
- Precision= 0.6658206
- Recall= 0.3919323
- F-Measure= 0.4934169
- Punto de corte= 0.4800

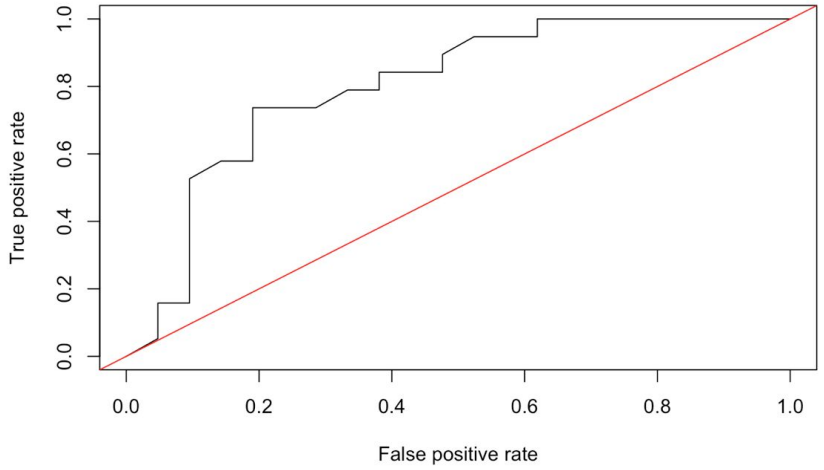
Se escoge el 0.4800 como punto de corte pues nos da una proporción de True Positives de 40.88%, una proporción de false positive de 6.3%, una proporción de true negatives de 93.7% y una proporción de false negatives de 59.11%. Idealmente, quisiéramos maximizar el True positive rate y minimizar el false negative rate; sin embargo, como se notó anteriormente, el set está desbalanceado pues más del 75% de los datos del set original tiene una clasificación original de 0 mientras que nuestro target es encontrar el valor 1 (label para los positives).

Modelo Naive Bayes

El tercer modelo que se aplicó fue el Naive Bayes. Este modelo asume independencia entre las variables predictoras y una distribución Gaussiana de éstas (lo cual definitivamente es un supuesto muy fuerte). Este modelo no requiere parámetros.

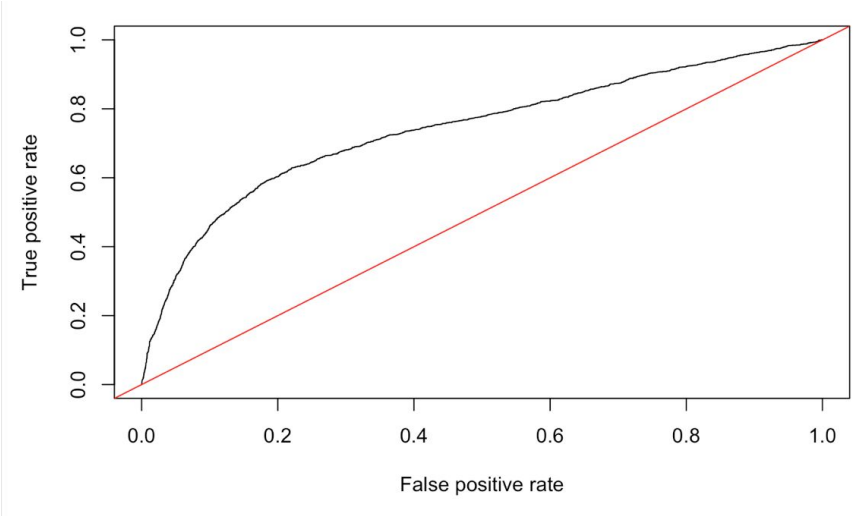
Desempeño con el set de validación

El área bajo la curva con el set de validación fue de 0.7273258, lo cual es menor que para los otros modelos (árbol y random forest).



El modelo naive bayes tiene una accuracy de 0.596 para el set de validación, lo cual es muy bajo y muy cercano a tirar una moneda y clasificar una observación según la cara que salga en la moneda. Además se puede ver que para un false positive rate menor a 0.1 la curva ROC está muy cerca de una línea de 45º, es decir, las clasificación se hacen casi como si se lanzará una moneda.

Desempeño con el set de pruebas



Ya con set de pruebas, el desempeño mejora y el área bajo la curva sube a 0.741421.

Matriz de confusión

	Reference	
Prediction	**0**	**1**
0	3963	494
1	3007	1541

Las medidas de desempeño son las siguientes:

- Accuracy = 0.6100468
- Precision= 0.3348817
- Recall = 0.7539841
- F-Measure= 0.463777

La accuracy del modelo es mucho menor que para los demás modelos, pues tanto en el árbol de decisión como en el modelo random forest la accuracy es mayor a 0.80. Curiosamente, la F-measure es mayor que para el árbol (que es 0.4532086). El recall es mayor que para los demás modelos.

Red Neuronal

Para poder aplicar la red neuronal el primer paso fue escalar las variables numéricas continuas, para eso se usó la función scale de R Studio: a cada observación numérica se le restó la media de su columna (es decir, la media de la variable) y después se divide entre la desviación estándar de la variable (columna). Para escalar el set de pruebas y el set de validación se usa la media y desviación estándar del set de entrenamiento.

Por la rapidez, se decidió correr la red neuronal usando Tensor Flow en Python con los siguientes parámetros: learning rate de 0.001, 20 training epochs, 10000 de batch size, 20 neuronas en la primera capa oculta, 20 neuronas en la segunda capa oculta, 23 neuronas input y 1 neurona en la capa de output (pues la clasificación es binaria). La red corrió rápidamente pero al momento de probar el desempeño con el set de pruebas se tuvieron resultados muy raros: accuracy de 1 (100%), Precision de 0, Recall de 0, y F-Measure de 0. La red neuronal era bastante compleja (por el número de capas ocultas y el número de neuronas en cada una). Para hacer el proceso más "transparente" se volvió a correr la red con una estructura diferente en R Studio.

Red Neuronal con 2 capas ocultas, 12 neuronas en cada una, función de activación logística y algoritmo backpropagation

Se intentó entrenar la red neuronal con el set de entrenamiento escalado (las variables categóricas no se escalaron), un threshold de 0.01, con 10000 iteraciones y corriendo la red 10 veces. En ninguna de las 10 repeticiones la red logró converger y tampoco pudo encontrar los pesos. El mínimo threshold era de 100.9681235, lo cual no tenía ningún sentido.

Red Neuronal con 2 capas ocultas, 12 neuronas en cada una, función de activación logística

Ya que si se usaba el algoritmo de backpropagation, el threshold era mayor a 100, se corrió la red con la función de activación logística pero sin el algoritmo de backpropagation. Asimismo, se decidió omitir el parámetro de learning rate. Se decidió ser "menos estrictos" y establecer un threshold de 0.05 y aumentar el número de iteraciones a 200,000. La red se tardó más de 12 horas en correr y no logró converger. El mínimo threshold fue de 0.562, pero la red no pudo encontrar los pesos.

Red Neuronal con 3 capas ocultas, 15 neuronas en cada una, función de activación logística

Se tenía la hipótesis que quizás si se aumentaba la complejidad del modelo el modelo lograría converger y encontrar los pesos adecuados para las variables predictoras. Se decidió agregar otra capa oculta y aumentar el número de neuronas en cada una. Se establecieron 200,000 iteraciones, threshold de 0.05 y 5 repeticiones. La red tardó casi un día en correr, el mínimo threshold fue de 0.7966 pero la red no pudo converger ni encontrar los pesos.

Resultados

Se aplicaron 4 distintos algoritmos de clasificación para poder encontrar el mejor modelo para saber si un determinado cliente de un banco en Taiwán iba a pagar su tarjeta de crédito el siguiente mes o no. Las medidas de desempeño con el set de pruebas fueron las siguientes:

Medida	**Árbol de decisión**	**Random Forest**	**Naive Bayes**	**Red Neuronal**
Área bajo la curva	0.7644124	0.7753478	0.741421	NA
Accuracy	81.78%	82%	61%	NA
Precision	68.90%	66.6%	33.5%	NA
Recall	33.76%	39.2%	75.4%	NA
F- Measure	0.4532086	0.4934169	0.463777	NA
Punto de corte	0.4863387	0.4800	-	NA

El modelo de clasificación fue el modelo de ensamble random forest con 300 árboles y node size de 5 observaciones. Este algoritmo es bastante útil pues añade diversidad y aleatoriedad al árbol de decisión. En todos los modelos fue claro que los algoritmos eran mejores clasificando las observaciones como categoría 0 (no cae en default) y esto hace sentido pues casi el 75% de los datos en el set original tenían esta clasificación. Aunque en este estudio se buscaba maximizar la proporción de true positives (las observaciones que el modelo clasifica como 1, es decir que sí van a caer en default el siguiente mes, fueran correctas) esto no fue posible. Más bien, los modelos eran mejores en la proporción de True Negatives, es decir, clasificar observaciones con categoría 0.

Conclusiones y estudios futuros

El objetivo de este estudio era encontrar el mejor modelo de clasificación para un Banco que busca saber si un determinado cliente suyo va a caer en default en su pago de tarjeta de crédito o no. Esto con el fin de poder tomar decisiones objetivas y que le permitirían al banco tomar precauciones respecto a sus clientes y cuidar su rentabilidad.

El mejor modelo de clasificación fue un modelo random forest con 300 árboles y node size de 10. Al igual que los otros modelos, este modelo clasificaba mejor las observaciones con categoría original 0. Se concluye que esto se debe al desbalance que tiene el set de datos original: con 75% de los datos con categoría 0 y solo el 25% con categoría 1. Asimismo, en el análisis exploratorio se pudo observar que en todas las variables numéricas continuas había más de 2000 observaciones con valores outliers, y que el rango (diferencia entre valor máximo y mínimo) para los valores de la variable era muy alto y que esto provocaba una desviación estándar muy alta e incluso mayor a la media.

Se concluye que el gran parte del desempeño de un modelo proviene de qué tan limpios y proporcionados estén los datos. Como estudio futuro se propone hacer un análisis por separado de las observaciones outliers y hacer simulación para poder lograr un modelo balanceado según la variable target.

Referencias

Notas de clase de Liliana Millán.