

# Modelo de predicción en siniestros de clientes SGR

*Joaquín Cavieres G.*

## 1. INFORMACIÓN DISPONIBLE

Para modelar y predecir la siniestralidad de clientes se cuenta con la siguiente información:

- a) Datos cuadratura registro de operaciones: Información relacionada a clientes junto a sus respectivas variables de interés en base de datos de AVLA
- b) Información Equifax consolidada: Información relacionada a clientes en base de Equifax (datos comerciales de clientes)
- c) Datos cuadratura mutuos comerciales: Información relacionada a clientes y respectivas variables de interés en base de datos de AVLA
- d) Datos consolidados Servicios Impuestos Internos (S.I.I): Información relacionada a clientes de Servicio Impuestos Internos.

## 2. CRUCE DE INFORMACIÓN

Para una mejor comprensión lectora de aquí en adelante se utilizan los siguientes nombres para las bases de datos:

- Datos cuadratura registro de operaciones + Datos cuadratura mutuos comerciales + Datos consolidados S.I.I  $\Rightarrow$  *data\_cuadra*
- Información Equifax consolidado  $\Rightarrow$  *data\_equi*

La información disponible en ‘data\_cuadra’ contiene inicialmente un total de 18331 registros de clientes en un rango de fecha desde 14 junio 2019 a 24 junio 2019 y con 5589 rut únicos. Posteriormente, esta base se filtra por ‘Fondo Póliza de Seguros First Aval’ y ‘Fondo Póliza de Seguro Avla’, para quedar con un total de 16171 registros.

La información proveniente de ‘data\_mutuos’ contiene inicialmente un total de 147 registros de clientes en un rango de fecha desde febrero 2018 hasta mayo de 2019. La información de datos de Servicios Impuestos Internos contenía 53393 registros y 6 variables como columnas informativas.

En primera instancia se unificaron, en función de las mismas variables en ambos conjuntos de datos, las bases de ‘data\_cuadra’ con ‘data\_mutuos’. Posteriormente, luego de esta unificación, la base se cruza con la base de datos de Servicios Impuestos Internos (con un total de registros iniciales de 53393). Por lo anterior es que crea una nueva base (con el mismo nombre ‘data\_cuadra’) cruzando la información entre ‘data\_cuadra’ y ‘data\_sii’, quedando una base de datos con 15813 registros.

La información de la base ‘data\_equi’ contiene inicialmente un total de 2821107 registros. Como esta base puede contener información semanal completa o información semanal parcial, entonces se genera un único registro mensual para cada cliente considerando un diferencial de -30 días y +30 días sobre el último registro. Luego de esto obtenemos una observación por mes y la base pasa a tener 1682 registros.

Adicionalmente, se integró una nueva información relacionada con los clientes que son normalizados. Estos clientes son aquellos que dado su mal comportamiento crediticio han tenido múltiples operaciones tratando de mejorar su situación financiera, pero, dado el conocimiento del negocio, en su gran mayoría, generalmente pasan a ser clientes siniestrados. Por lo anterior es que utilizará como complemento de la variable a modelar.

Para generalizar la variable a modelar se consideró a la columna ‘ESTADO\_CERTIFICADO’ de la base ‘data\_cuadra’ para el modelo de predicción. Esta nueva variable tiene por nombre **SINIESTRADO** y cuenta con sólo dos posibles resultados:  $\Rightarrow$  **Si/No**. Además, dada la información disponible de los clientes normalizados, se construye un nuevo modelo considerando que: Si algún rut de la base comercial aparece al

menos una vez en la base de normalización, entonces, ese cliente será marcado como SINIESTRADO en su certificado más reciente.

Por lo anterior el desglose de la primera marca como SINIESTRADO es:

SINIESTRADO	No	Si
Activo	✓	
Cancelado	✓	
Pagado ACh		✓
Siniestro parcial		✓

y la segunda marca para SINIESTRADO2 queda como:

SINIESTRADO2	No	Si
Activo	✓	
Cancelado	✓	
Pagado ACh		✓
Siniestro parcial		✓
Normalizado		✓

De lo anterior se puede decir que los SINIESTRADOS / NO SINIESTRADOS para las marcas propuestas son:

Marcas	No	Si
SINIESTRADO	1175	186
SINIESTRADO2	1053	308

## 2.1. PROCESAMIENTO ‘data\_cuadra’ y ‘data\_equi’

Primero se cuadró el campo RUT\_CLIENTE en ‘data\_cuadra’ ya que, por ejemplo, un registro estaba especificado por ‘CL995355000’, por lo tanto se traspasa a ‘995355000’. Para ‘data\_equi’ se hace el mismo tratamiento ya que en algunos registros el campo RUT es ‘0690721007’ para pasar a ‘690721007’. Este mismo procedimiento se realizó en ‘data\_sii’ para configurar el campo RUT con la finalidad de cruzar con ‘data\_cuadra’ y así incorporar la información disponible del Servicio de Impuestos Internos.

En el campo RUT de ‘data\_equi’ también vienen registros del estilo ‘0000001031’ los que son eliminados de la base de datos ya que no coincidirían con los que quieren ser pareados de ‘data\_cuadra’.

Posteriormente, para hacer coincidir la información de ‘data\_cuadra’ y ‘data\_equi’, también se integra el campo fecha (que por nombre aparece como FECHA\_EMISION y date respectivamente). Se hizo una conversión de estos campos para coincidir formatos y el cruce final de información.

Ya que la finalidad es cruzar la información de RUT, AÑO, MES en ambos conjuntos de datos, y por otro lado ‘data\_equi’ puede contener hasta 4 registros en un mes para un mismo cliente, se trabajó con la menor diferencia negativa con un máximo de -30 días, o por el contrario, usamos la menor diferencia positiva con un máximo de 30 días.

Todo lo anterior se resume en el siguiente esquema:

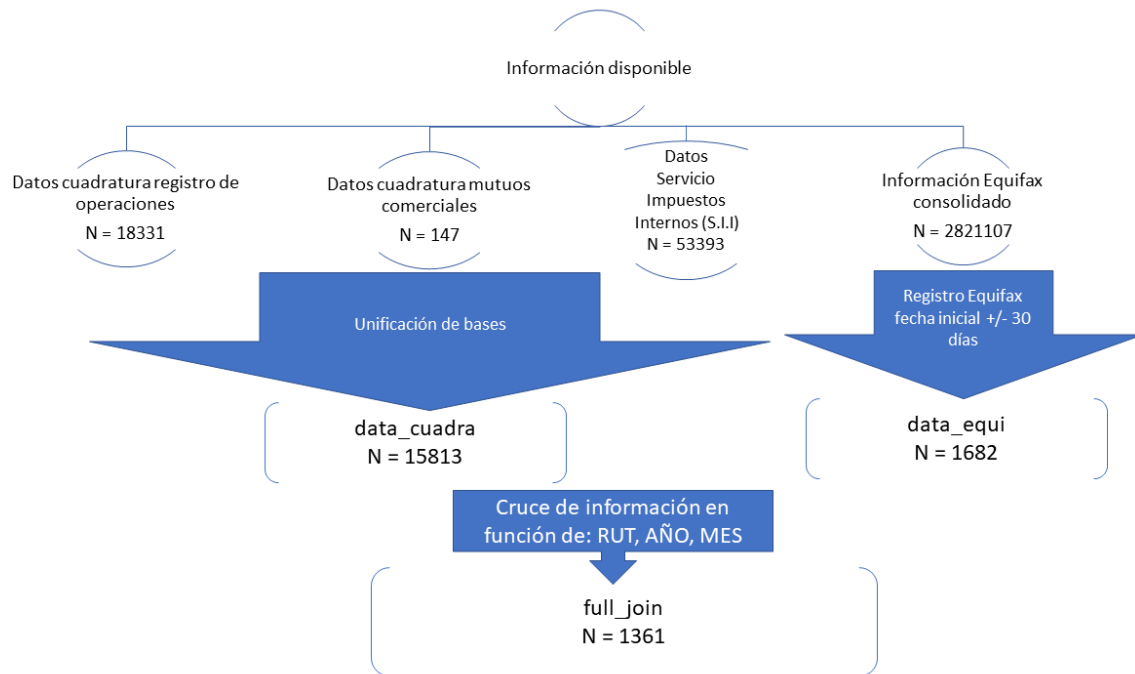


Figura 1: Procesamiento de información SGR

## 2.2. ANÁLISIS DESCRIPTIVO PREVIO

A continuación se presentan algunos análisis descriptivos previos de las variables numéricas relacionadas a 'data\_cuadra'.

### 2.2.1. Variables de interés

Se presenta la siguiente tabla con los estadísticos de resumen para las variables DIAS\_GRACIA, MONTO\_CERT\_PESOS Y NUM\_CUOTAS (Días de gracia para el pago del préstamo, monto del préstamo en pesos y número de cuotas para pago del préstamo respectivamente). Estas son las variables numéricas más representativas del conjunto de datos ya que las demás no presentan gran variación (en términos de variabilidad). La variable MONTO\_CERT\_PESOS es la que tiene una mayor variabilidad entre los registros de 'data\_cuadra' seguido por NUM\_CUOTAS y DIAS\_GRACIA. La mediana de los registros para MONTO\_CERT\_PESOS está entre los \$50000000 pesos y un máximo de \$119750499 pesos. Para el número de cuotas el máximo es de 144 y el mínimo de 1. Finalmente para DIAS\_GRACIA la mediana se concentra en los 75 días días.

	DIAS_GRACIA	MONTO_CERT_PESOS	NUM_CUOTAS
<b>Mean</b>	123.06	116822552.35	23.77
<b>Std.Dev</b>	152.50	183923305.10	29.55
<b>Min</b>	0.00	100000.00	1.00
<b>Q1</b>	34.00	22425150.00	1.00
<b>Median</b>	75.50	50000000.00	9.00
<b>Q3</b>	153.00	119750499.00	36.00
<b>Max</b>	1634.00	1668154163.00	144.00
<b>MAD</b>	65.98	51504962.09	11.86
<b>IQR</b>	119.00	97325349.00	35.00
<b>CV</b>	1.24	1.57	1.24
<b>Skewness</b>	4.02	3.47	1.50
<b>SE.Skewness</b>	0.07	0.07	0.07
<b>Kurtosis</b>	25.75	16.17	2.17
<b>N.Valid</b>	1361.00	1361.00	1361.00
<b>Pct.Valid</b>	100.00	100.00	100.00

La siguiente tabla descriptiva esta relacionada al conjunto de datos ‘data\_equi’. Principalmente se describen las variables que podrían tener un mayor impacto en el comportamiento de los clientes y el riesgo financiero que presentan. Los estadísticos de resumen muestran que la mediana se concentra cerca del 0 en la mayoría de los casos pero con medias variables. Si consideramos la media como un estadístico de medida central entonces para la variable Monto\_Mora\_Pesos presenta una promedio de \$6771010, mientras que la variable Monto\_Protestos\_Pesos tiene una media de \$46894. Las variables N\_Moras, N\_Multas y N\_Protestos presentan medias entre las 249, 69 y 298 respectivamente.

Error in where(obj\_name) : length(name) == 1 is not TRUE

	Monto_Mora_Pesos	Monto_Protestos_Pesos	N_Moras	N_Multas	N_Protestos
<b>Mean</b>	6771010.43	46894.37	2.16	0.15	1.50
<b>Std.Dev</b>	48213118.26	865864.37	11.71	2.03	10.96
<b>Min</b>	0.00	0.00	0.00	0.00	0.00
<b>Q1</b>	0.00	0.00	0.00	0.00	0.00
<b>Median</b>	0.00	0.00	0.00	0.00	0.00
<b>Q3</b>	83600.00	0.00	1.00	0.00	0.00
<b>Max</b>	850053262.00	27204802.00	249.00	69.00	298.00
<b>MAD</b>	0.00	0.00	0.00	0.00	0.00
<b>IQR</b>	83600.00	0.00	1.00	0.00	0.00
<b>CV</b>	7.12	18.46	5.43	13.51	7.33
<b>Skewness</b>	12.25	26.30	14.70	29.49	19.22
<b>SE.Skewness</b>	0.07	0.07	0.07	0.07	0.07
<b>Kurtosis</b>	175.00	763.81	256.28	969.91	458.84
<b>N.Valid</b>	1361.00	1361.00	1361.00	1361.00	1361.00
<b>Pct.Valid</b>	100.00	100.00	100.00	100.00	100.00

La siguiente figura muestra la correlación existente entre las variables de interés más representativas en los conjuntos de datos 'data\_cuadra' y 'data\_equi'. Para las variables de 'data\_cuadra' la correlación entre ellas no es alta, incluso teniendo valores negativos, lo que nos indica una linealidad inversa entre dichas variables (DIAS\_GRACIA y NUM\_CUOTAS). Lo anterior nos permite incorporarlas dentro del predictor evitando la multicolinealidad y sin perder la dimensionalidad de nuestros datos.

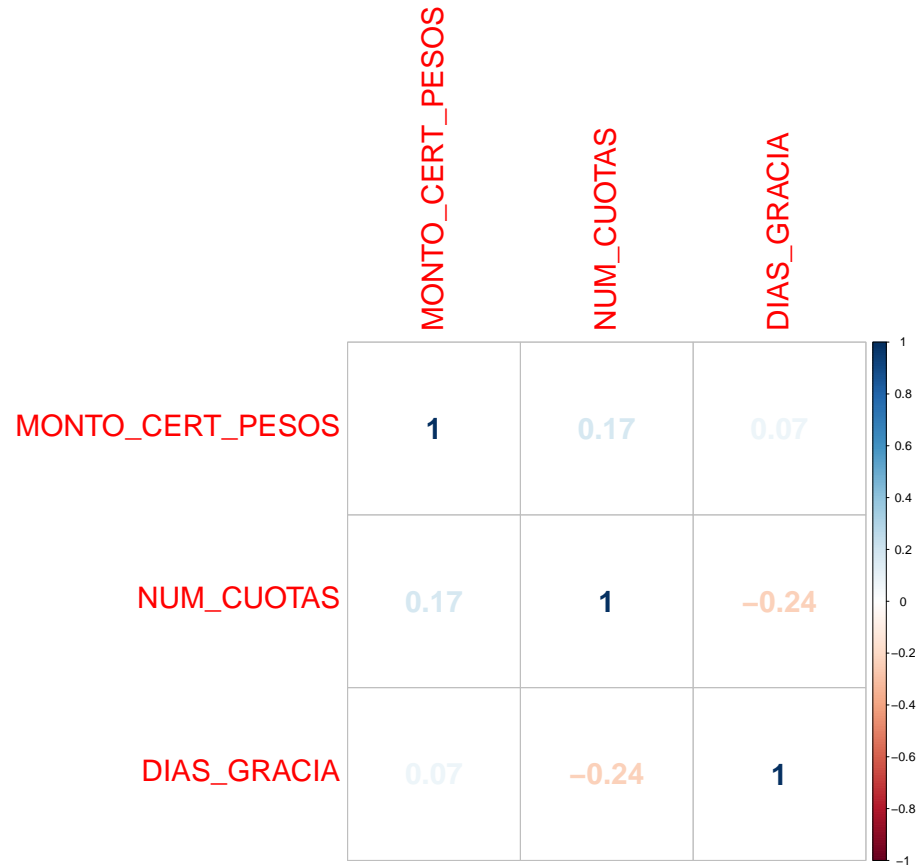


Figura 2: Correlación variables data cuadra

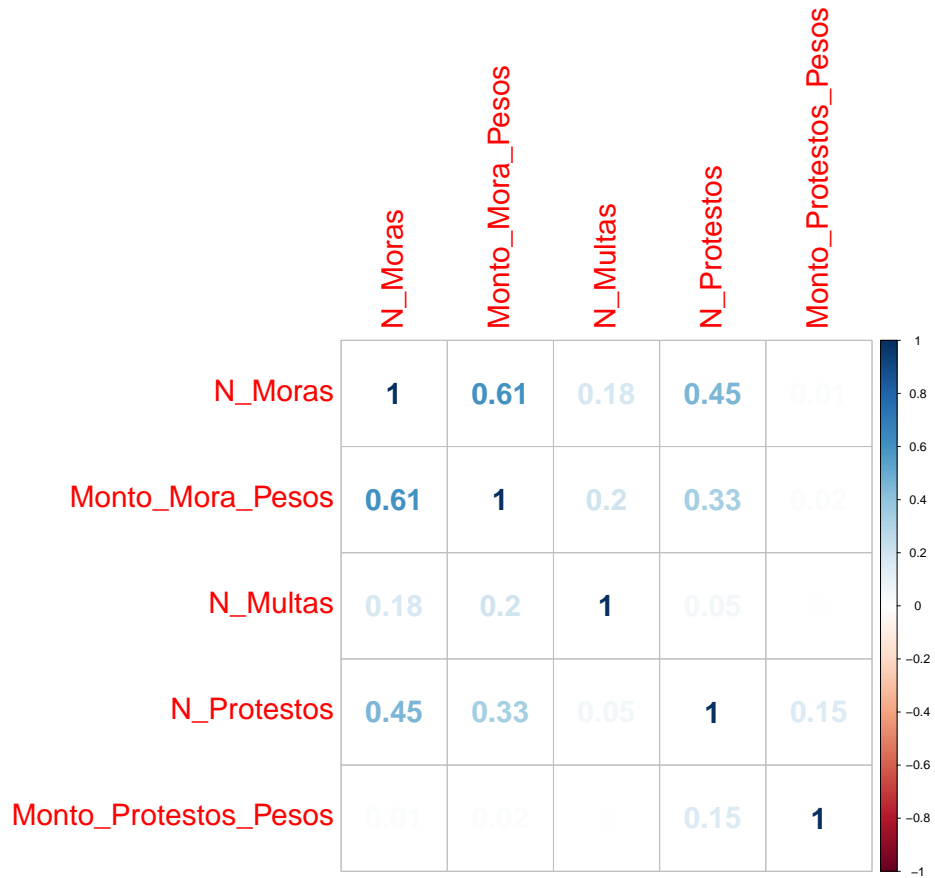


Figura 3: Correlación variables data equi

La correlación existente en algunas variables de ‘data\_equi’ alcanza en algunos casos un valor de 0.61 (MONTO\_MORA\_PESOS y N\_MORAS), pero no siendo un valor altamente significativo ( $<0.7$ ), por lo que no induciría en un problema de multicolinealidad dentro del predictor lineal. Sin embargo, esto no quiere decir que cada una por sí sola pueda explicar el comportamiento de SINIESTRADO o NO SINIESTRADO, la suma total de estas sí podría tener directa influencia sobre la probabilidad de que un cliente sea efectivamente SINIESTRADO, pero es necesario evaluarlas mediante los modelos de predicción que se propondrán más adelante.

### 2.2.2. Análisis gráfico sobre marca SINIESTRADO

La figura 4 muestra la proporción de SINIESTRADOS (Si/No) en la primera marca propuesta a modelar. Los clientes que efectivamente SINIESTRAN alcanzan el 13.7% versus el 86.3% de los que NO SINIESTRAN.

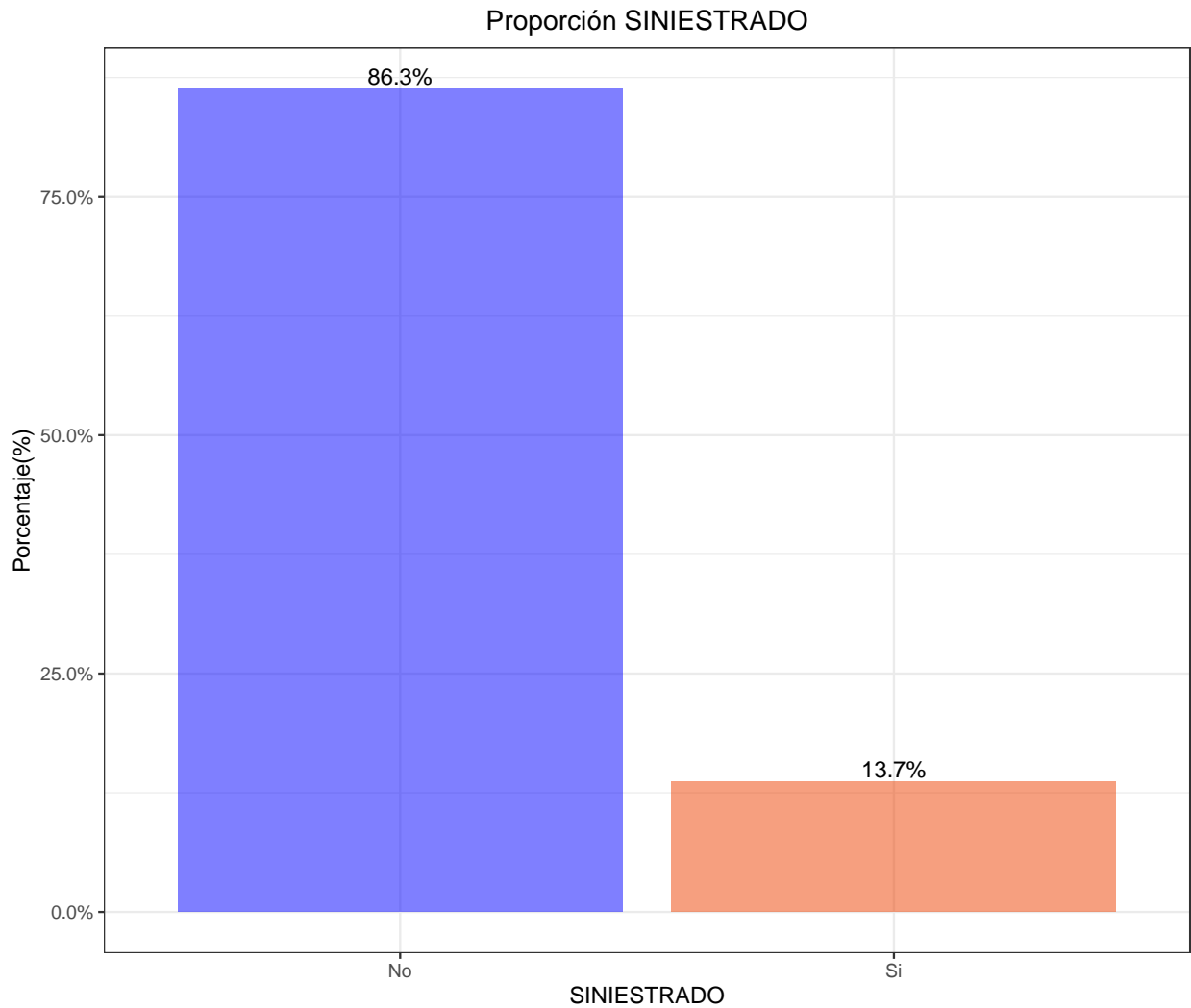


Figura 4

*Nota: Todos los siguientes gráficos están en función de la primera marca (SINIESTRADO). Más adelante se encuentra otra sección en donde se desarrollan los mismos análisis que se presentaron a continuación pero para la segunda marca (SINIESTRADO2)*

En la siguiente figura se puede apreciar visualmente que la distribución en escala logarítmica de los clientes que SINIESTRAN y NO SINIESTRAN es similar, por lo que la distribución de estas variables no tienen relación con que un cliente SINIESTRA o NO SINIESTRA (ver Anexo II), sin embargo esto no quiere decir que el que SINIESTRAR/NO SINIESTRAR sean dependientes uno del otro.

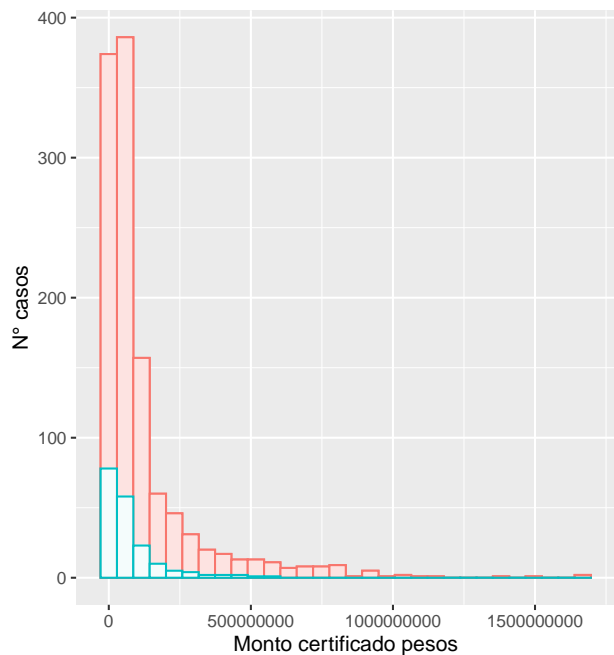


Figura 5

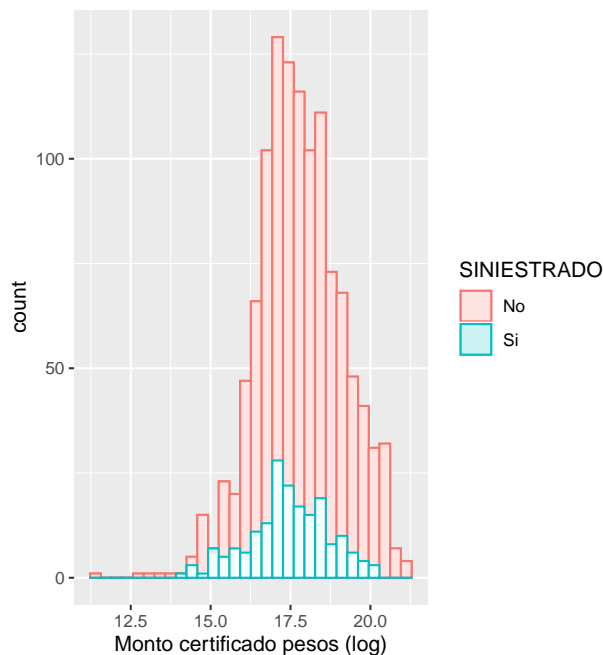


Figura 6

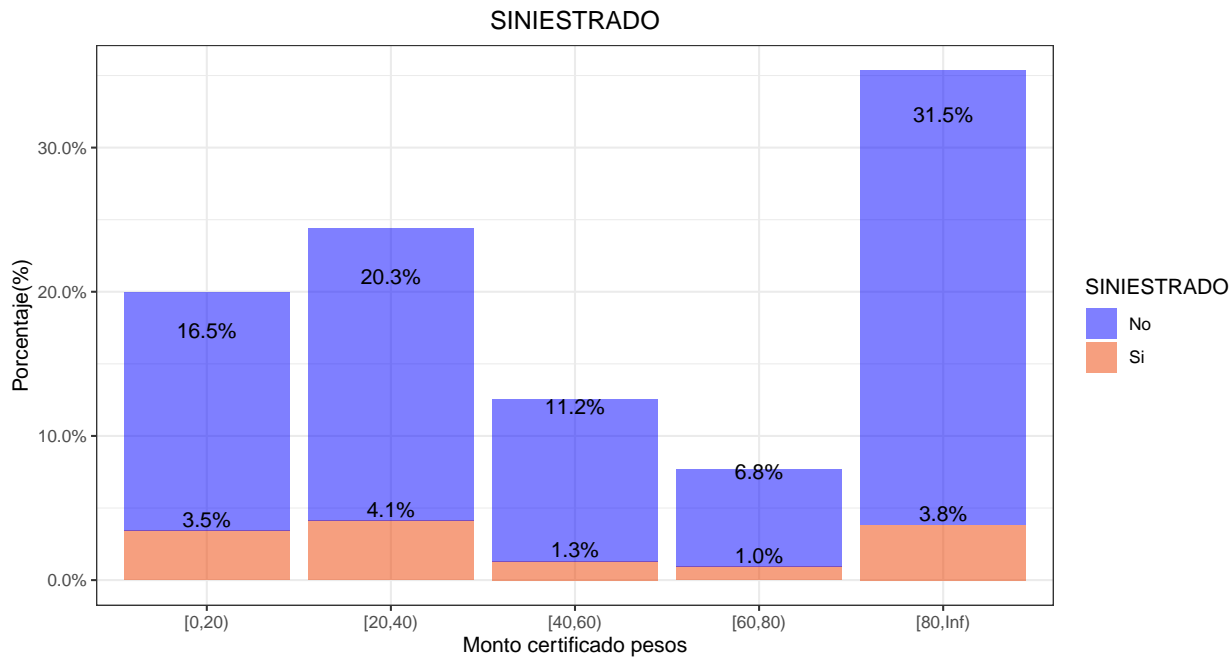


Figura 7

*Nota: El gráfico anterior (Monto de certificado), para representar de mejor forma la escala del eje x, el valor fue dividido por \$1.000.000*



El siguiente gráfico está relacionado al número de cuotas (NUM\_CUOTAS) en el cual un cliente pacta su contrato. Se puede apreciar que la mayoría de los SINIESTRADOS se concentran sobre las 20 cuotas, mientras que un pequeño porcentaje se encuentra entre 0 - 5 cuotas (2.7%), y otro porcentaje más bajo con cuotas entre 10 y 15 (0.8%)

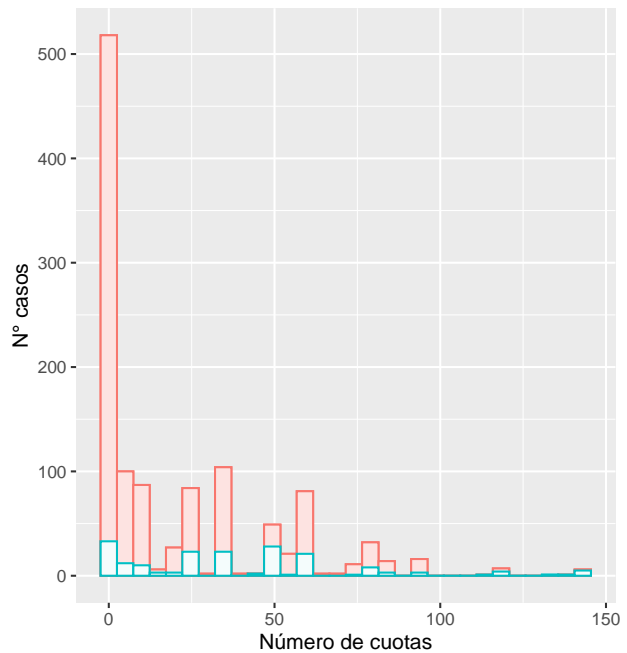


Figura 8

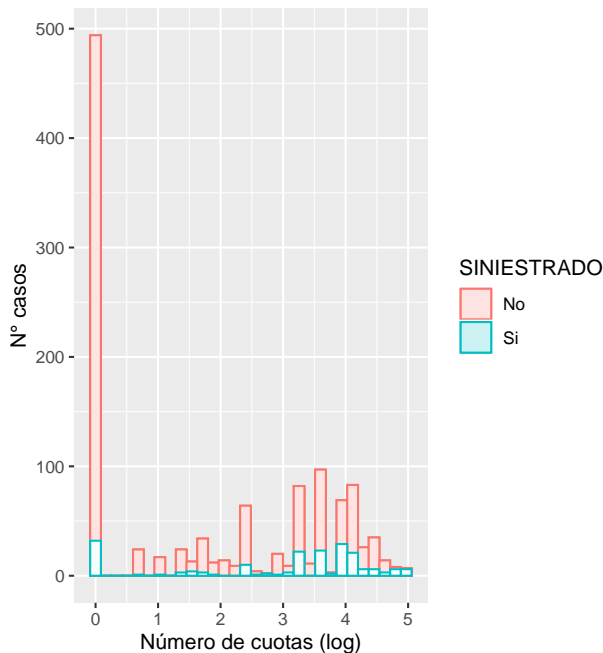


Figura 9

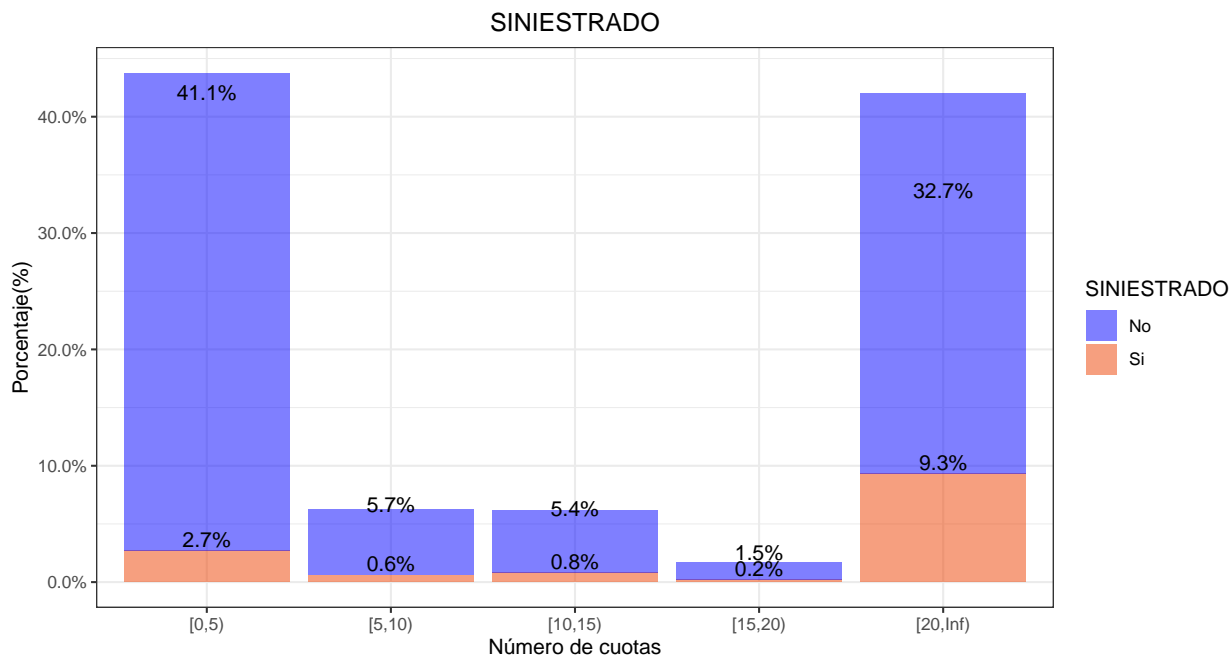


Figura 10

La variable DIAS\_GRACIA tiene el mismo comportamiento de las anteriores, pero teniendo una distribución más variable entre ellos. Los mayoría de los clientes SINISTRADOS se encuentran en el rango de los 0 - 40 días (5.2%), mientras que los clientes con días de gracia mayor a 160 siniestran 3.0%, y en un porcentaje menor los clientes entre 40 - 80 días de gracia (2.8%). En todos los rangos de días existen clientes SINIESTRADOS.

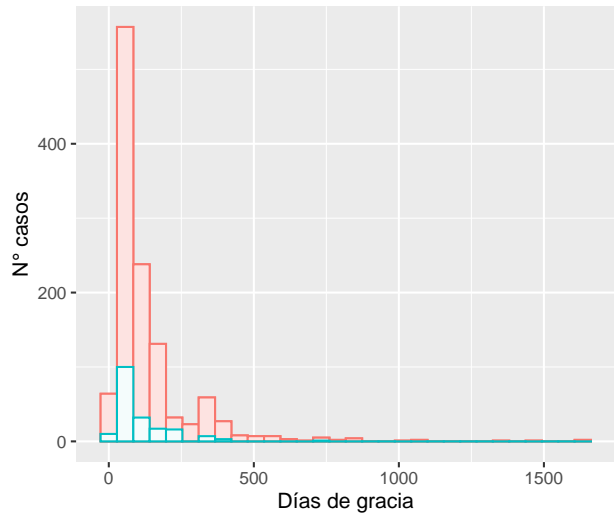


Figura 11

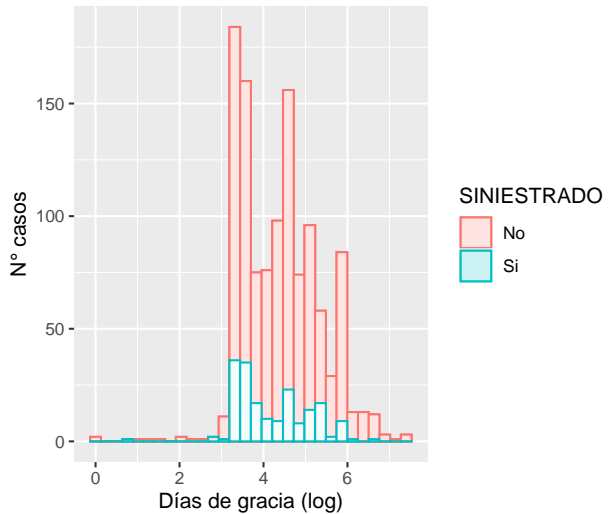


Figura 12

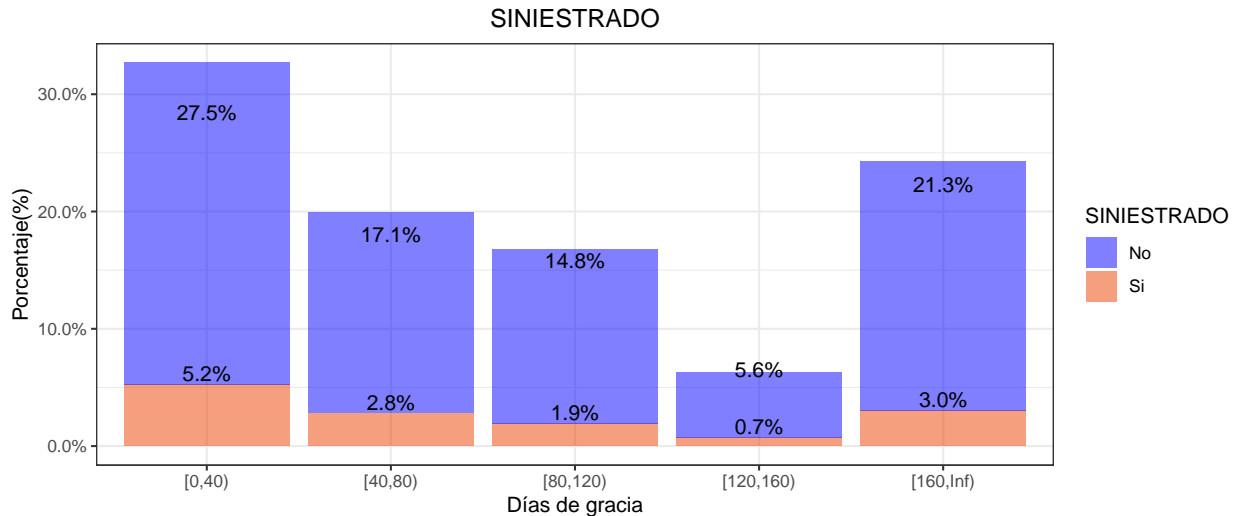
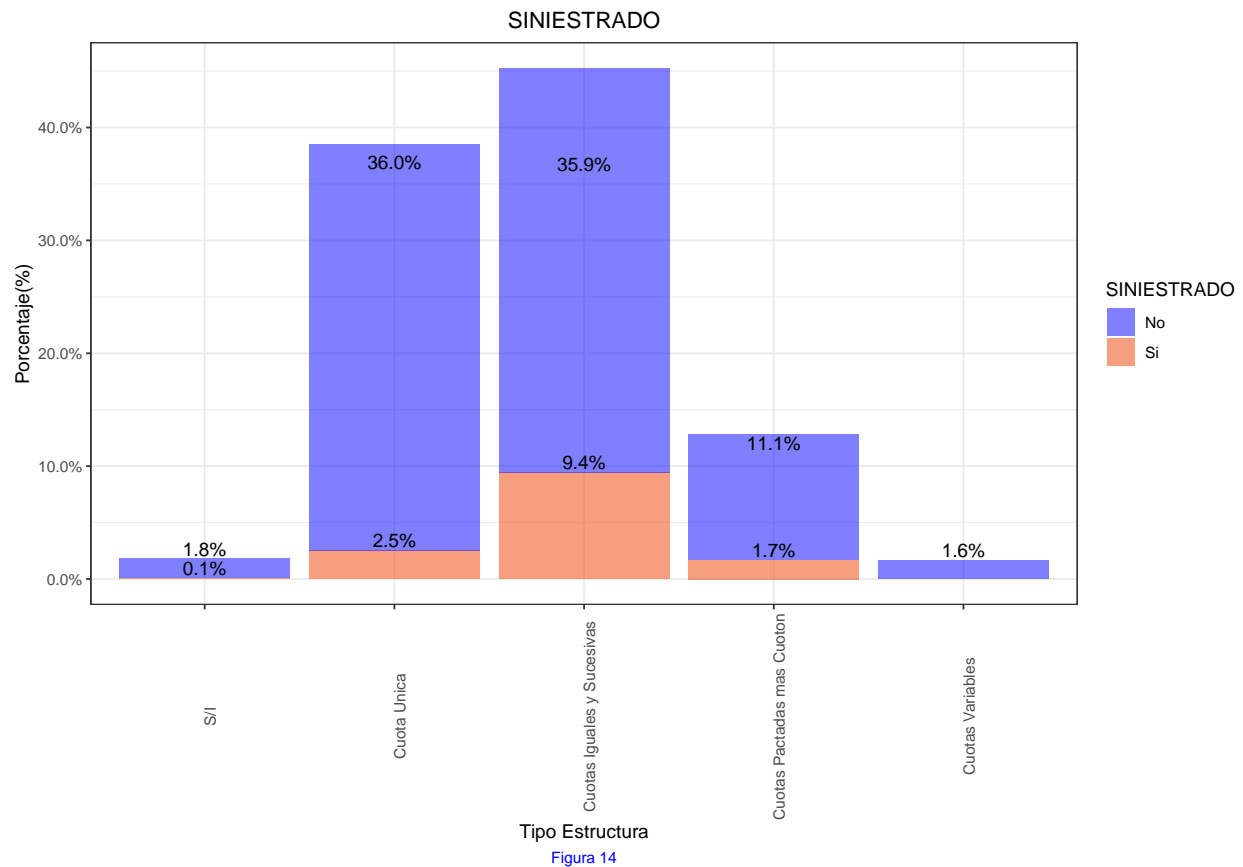


Figura 13

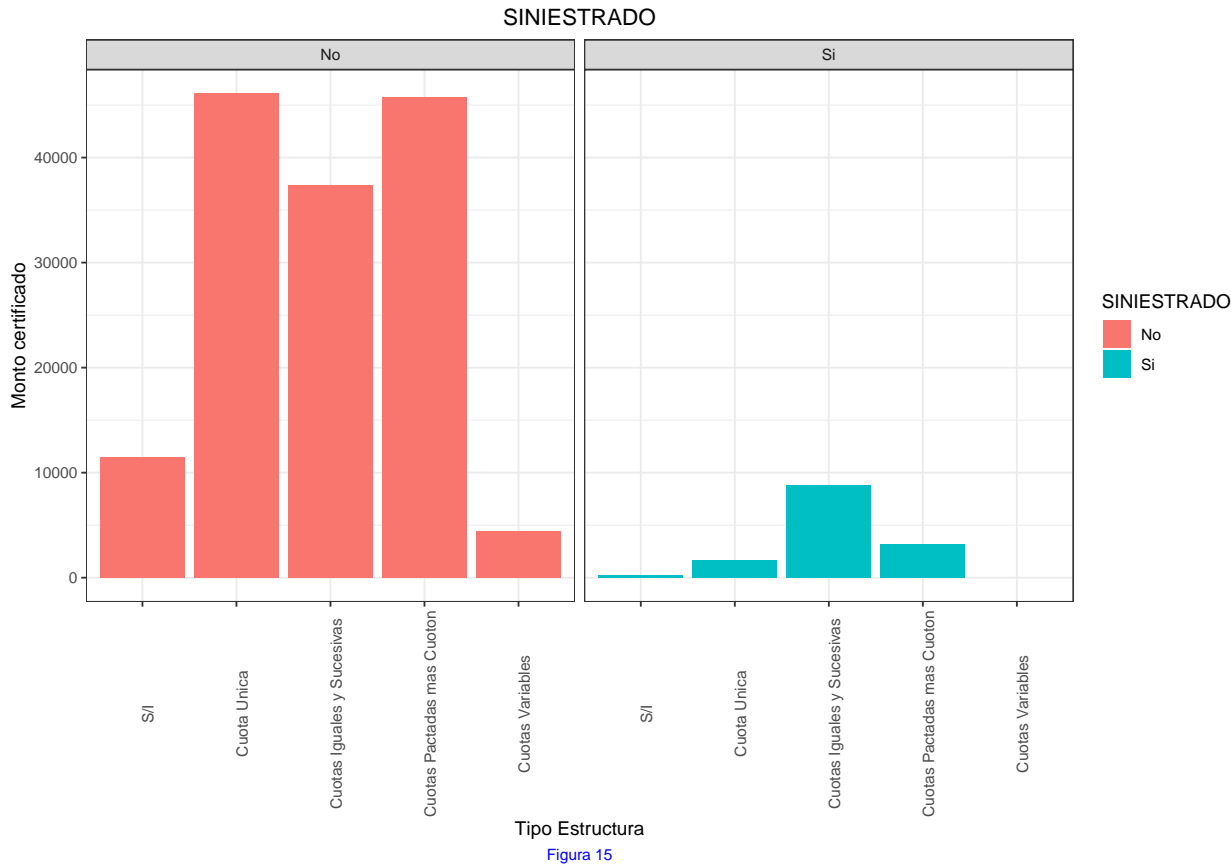
A continuación se presenta la variable categórica ‘TIPO\_ESTRUCTURA’ que se encuentra en ‘data\_cuadra’. Se aprecia que el mayor porcentaje de los clientes obtiene un contrato con ‘Cuota Iguales y Sucesivas’ (45.26%), seguido de ‘Cuota Única’ (38.50%) y de ‘Cuotas Pactadas más Cuotón’ (12.78%).

	Freq	%	% Cum.
<b>S/I</b>	25	1.84	1.84
<b>Cuota Unica</b>	524	38.50	40.34
<b>Cuotas Iguales y Sucesivas</b>	616	45.26	85.60
<b>Cuotas Pactadas mas Cuoton</b>	174	12.78	98.38
<b>Cuotas Variables</b>	22	1.62	100.00
<b>Total</b>	1361	100.00	100.00

La figura de ‘TIPO\_ESTRUCTURA’ versus SINIESTRADO muestra cómo la estructura del contrato también influye en el comportamiento de clientes que siniestran efectivamente. Un cliente que estructura su contrato con ‘Cuotas iguales y sucesivas’ y ‘Cuotas pactadas más Cuotón’ tiene una mayor proporción de SINIESTRADOS que las demás tipos de estructuras (9.4% y 1.7% respectivamente)



La relación entre TIPO\_ESTRUCTURA y MONTO\_CERT\_PESOS muestra que aquellos clientes que SINIESTRAN pactan un contrato en ‘Cuotas Iguales y Sucesivas’ y un monto del certificado en pesos promedio cercano a los \$10000000. Mientras que los clientes que pactan su contrato en ‘Cuota Única’ y cerca de los \$50000000 de pesos, son los que presentan una mayor proporción de NO SINIESTRAR.



*Nota: El gráfico anterior (Monto de certificado), para representar de mejor forma la escala del eje x, el valor fue dividido por \$1.000.000*

Los siguientes gráficos están relacionados a la variable SINIESTRADO y las variables del conjunto de datos 'data\_equi'. El gráfico de abajo nos muestra la clara diferencia entre los clientes que SINIESTRAN frente a la variable Monto\_Mora\_Pesos. Se ve que los clientes SINIESTRADOS se concentran preferentemente en el rango de los \$0 - \$200000 pesos con un 10.3%, mientras que los clientes con montos de moras cercanas a los \$800000 pesos SINIESTRAN cerca del 1.7%.

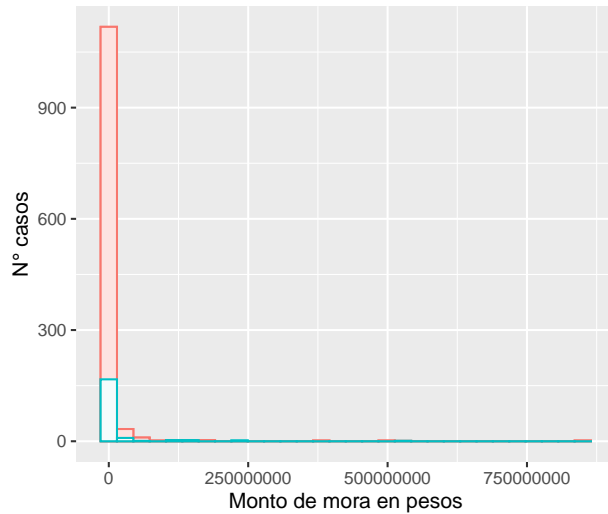


Figura 16

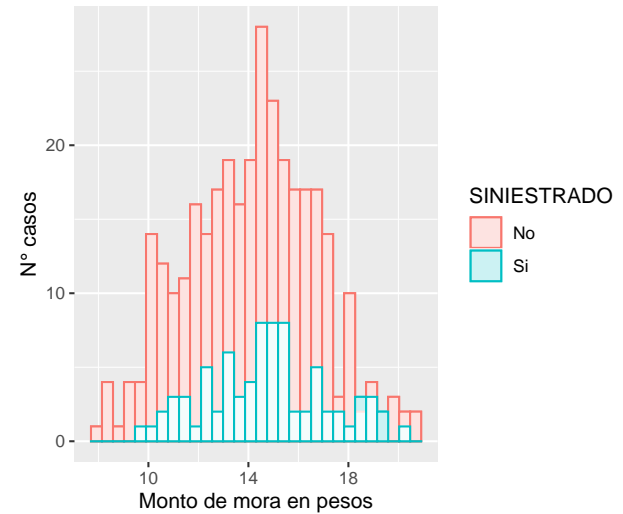


Figura 17

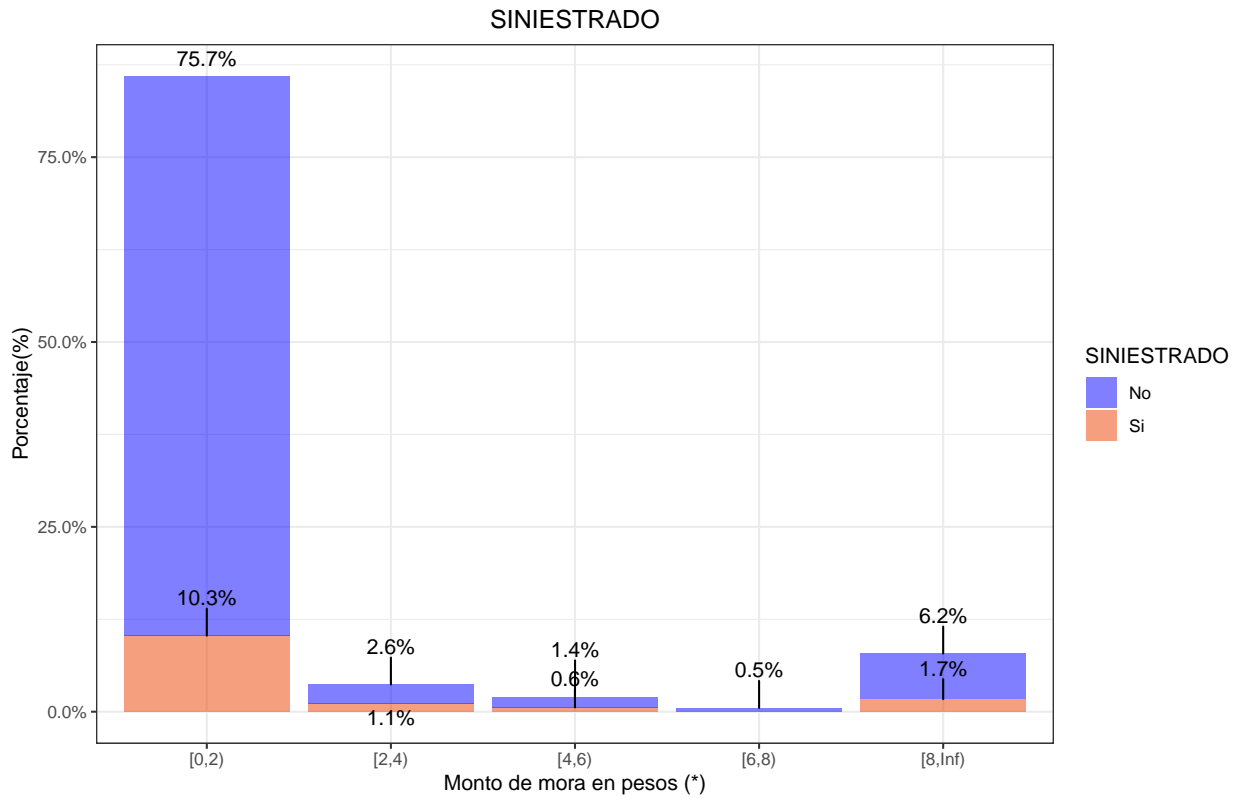


Figura 18

*Nota: El gráfico anterior (Monto de moras en pesos), para representar de mejor forma la escala del eje x, el valor fue dividido por \$1.000.000.*

El gráfico de N\_Moras también muestra homogeneidad entre los clientes SINIESTRADO / NO SINIESTRADO en términos distribucionales (escala original y escala logarítmica). Pero el gráfico siguiente muestra que el mayor porcentaje de los clientes “SINIESTRADOS” se encuentran en el rango de las 0 - 1 número de moras (7.9%), mientras el siguiente porcentaje más representativo de clientes SINIESTRADO se concentra entre número de moras mayores a 4 (2.9%)

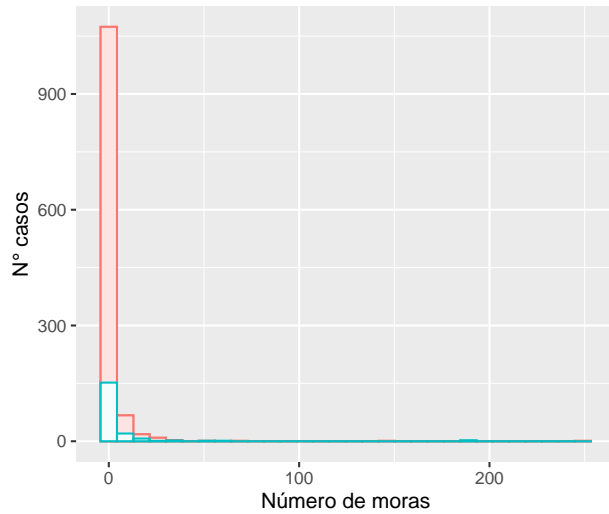


Figura 19

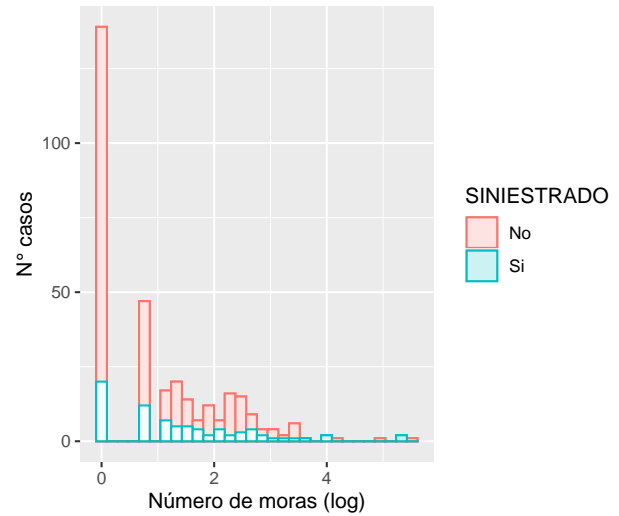


Figura 20

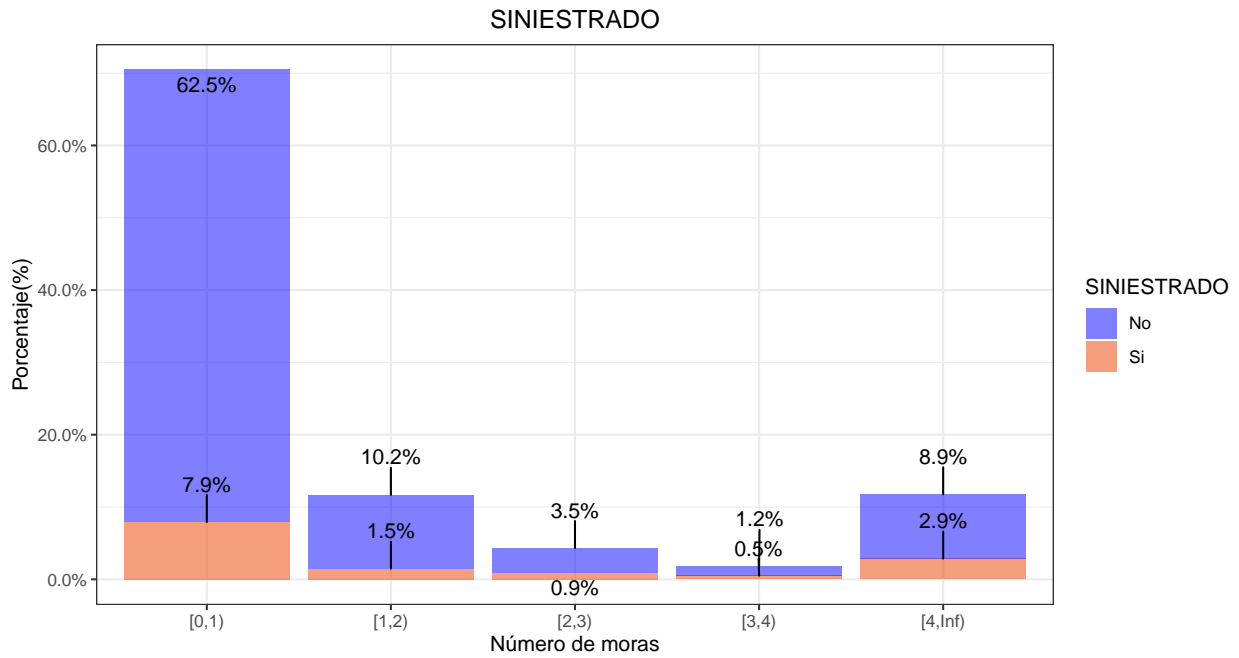


Figura 21

La variable N\_Multas también es homogénea entre clientes SINIESTRADO / NO SINIESTRADO. La mayor proporción de clientes SINIESTRADOS se concentra entre las 0 - 1 número de multas (12.6%) y los demás rangos de porcentajes SINIESTRADOS presentan valores muy bajos.

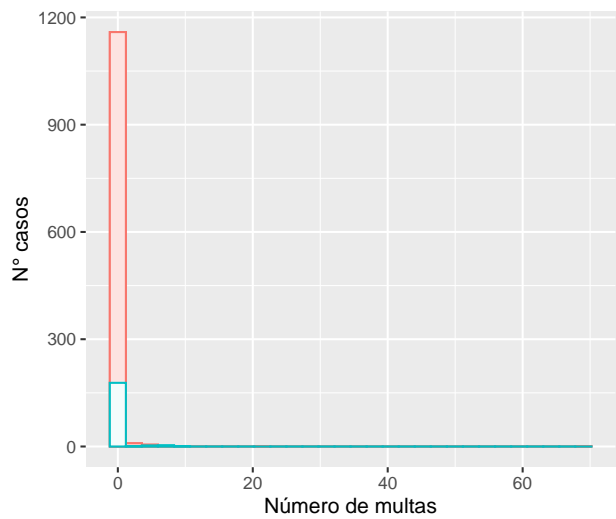


Figura 22

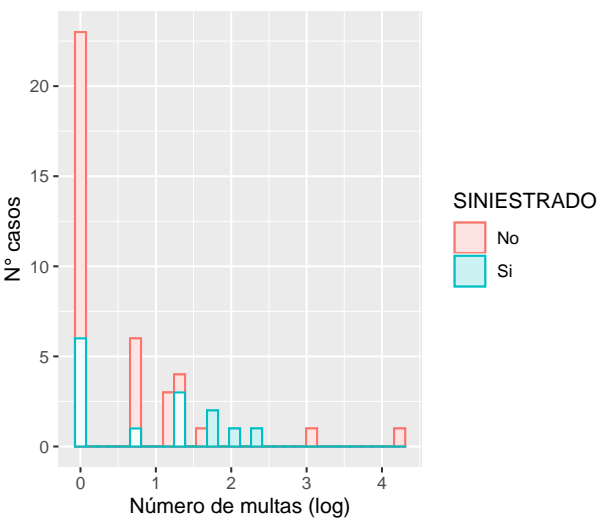


Figura 23

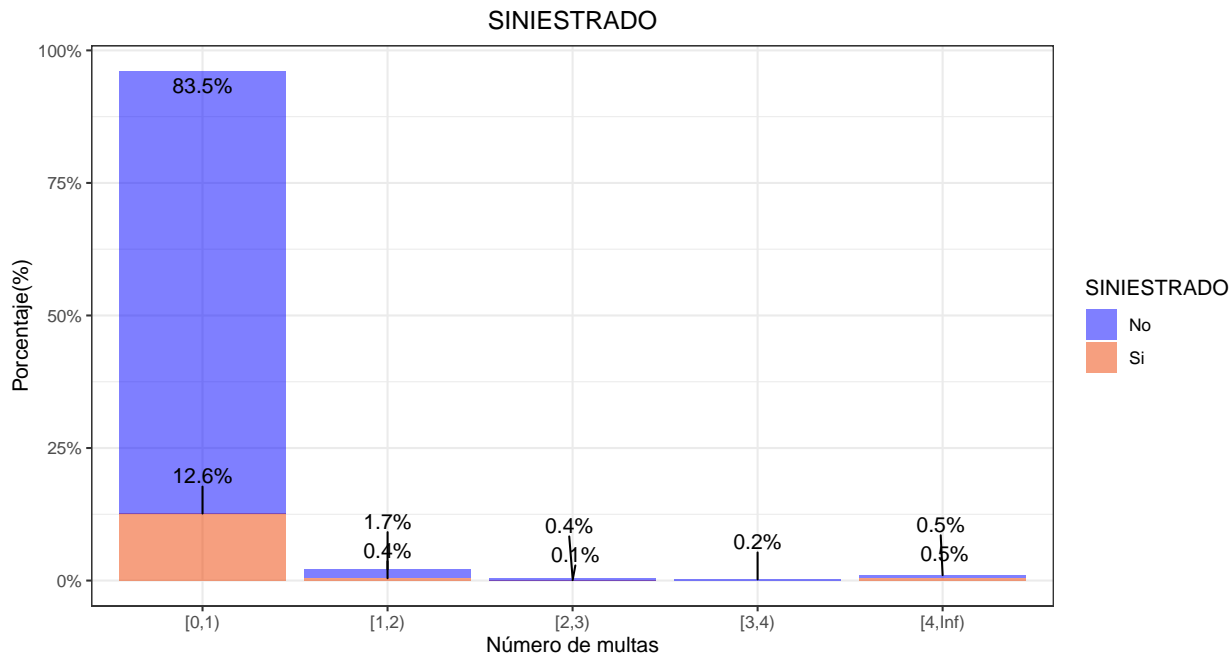


Figura 24

Como se vio anteriormente en los gráficos presentados, el mayor porcentaje de SINIESTRADO se concentra dentro del primer rango de distribución, en este caso dentro del 0 - 1 número de protestos (9.5%)

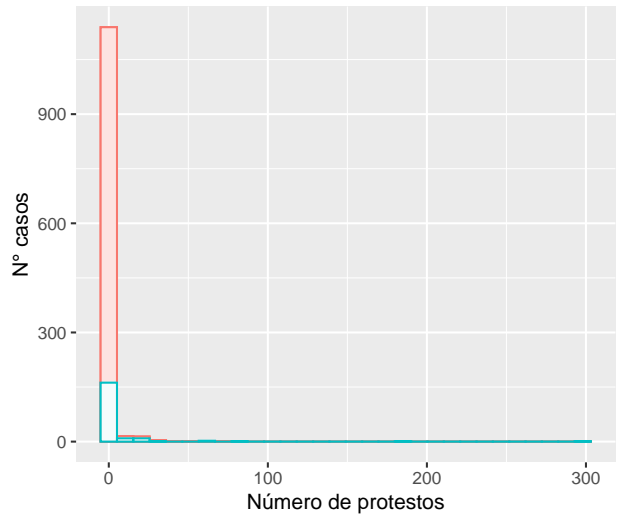


Figura 25

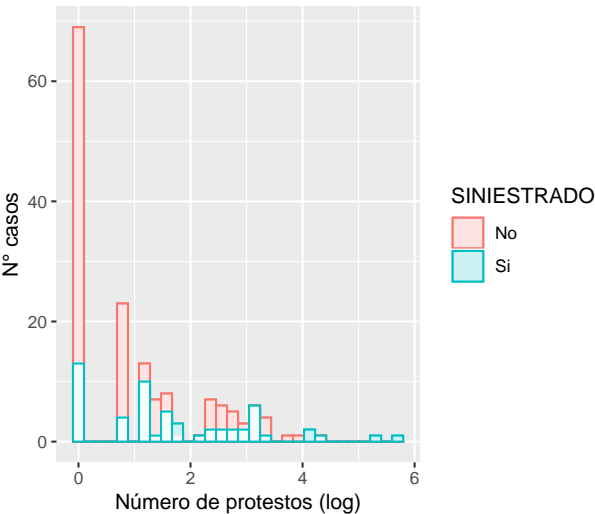


Figura 26

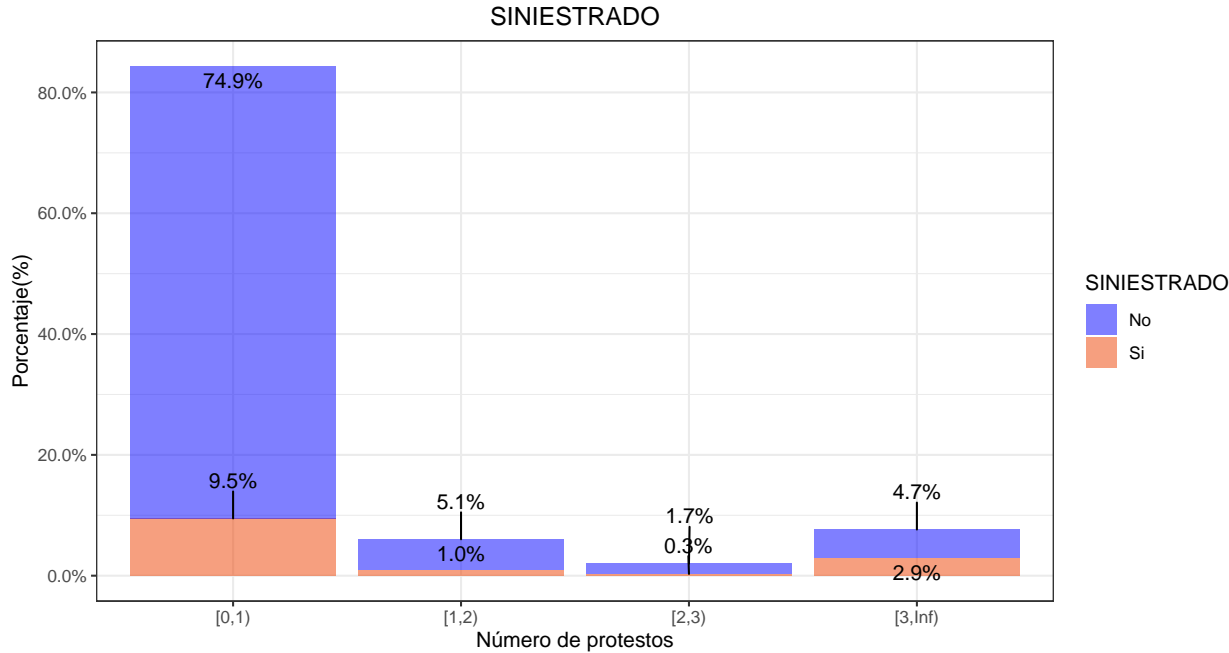


Figura 27



Finalmente, el gráfico a continuación presenta la variable Monto\_Protestos\_Pesos en función de los SINIESTRADOS / NO SINIESTRADOS. Como se vio anteriormente en los diferentes gráficos, el mayor porcentaje de SINIESTRADOS se concentra dentro del primer rango de distribución, y en escala logarítmica se puede apreciar que el Monto\_Protestos\_Pesos tiene un valor llamativo cercano a 15. Como se mostró en la tabla resumen de páginas anteriores, este monto puede ser uno de los pocos registros observados empíricamente (ya que contamos con 1348 observaciones con datos 0 y sólo 12 observaciones con observaciones efectivas mayores a 0). En relación a los clientes que SINIESTRAN, la mayor proporción se encuentra en el rango de 0 - 1, quienes son mayormente los clientes con 0 Monto de protestos en pesos con un 9.8% del total en ese rango.

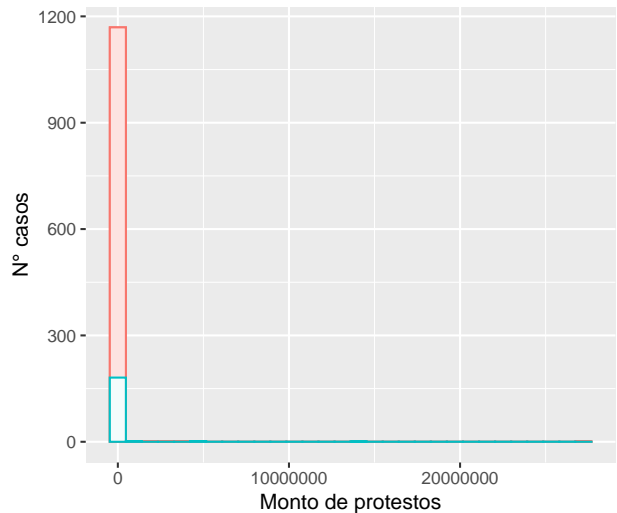


Figura 28

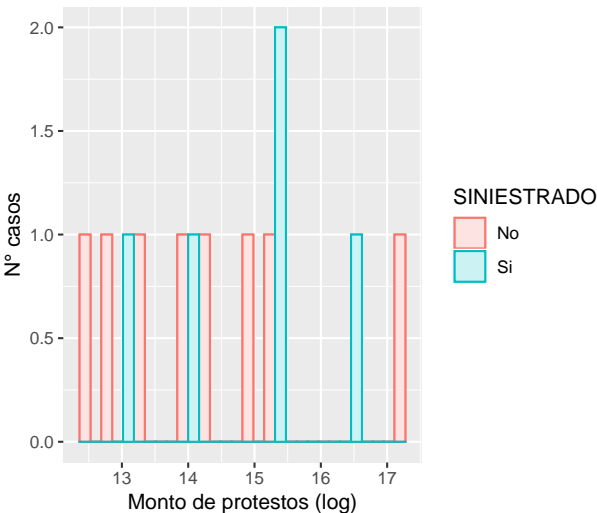


Figura 29

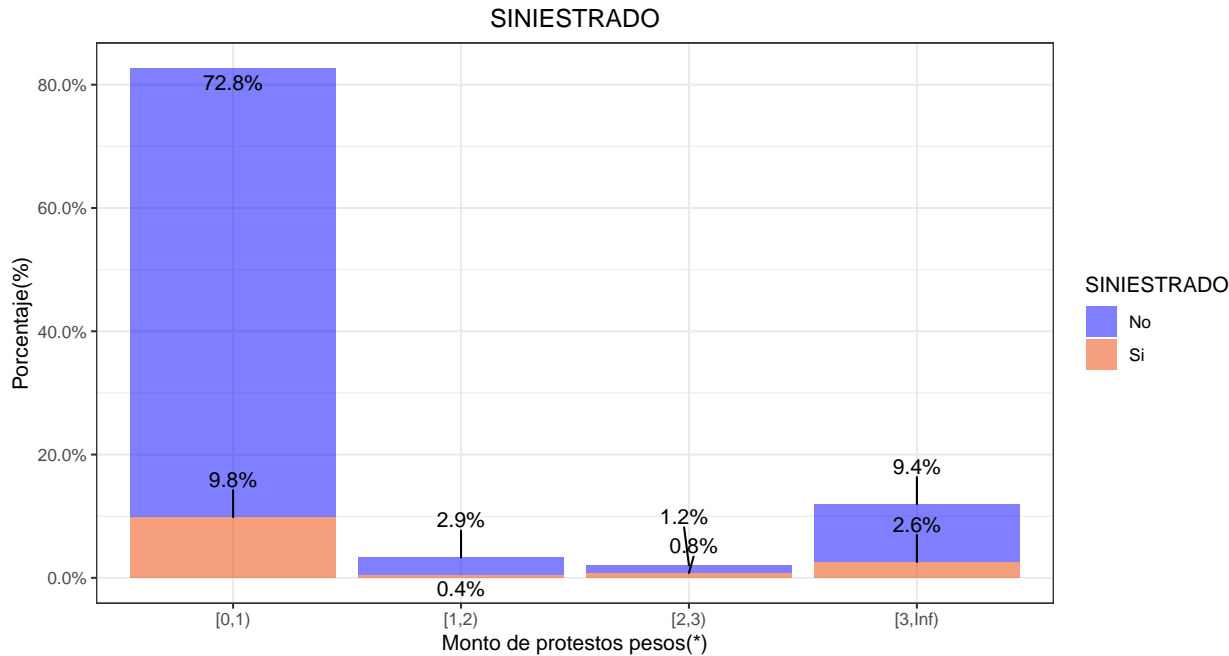


Figura 30

Nota: El gráfico anterior (Monto protestos en pesos), para representar de mejor forma la escala del eje x, el valor fue dividido por \$1.000.000.

Los siguientes gráficos buscan representar alguna relación existente entre una variable de ‘data\_cuadra’ con otro de ‘data\_equi’. Los clientes que presentan N\_Multas con valores altos y efectivamente SINIESTRAN, son aquellos clientes con contrato de ‘Cuotas Iguales y Sucesivas’, mientras que aquellos que tienen un número bajo de N\_Multas y también SINIESTRAN, tienen una estructura de tipo contrato de ‘Cuotas Pactadas más Cuoton’.

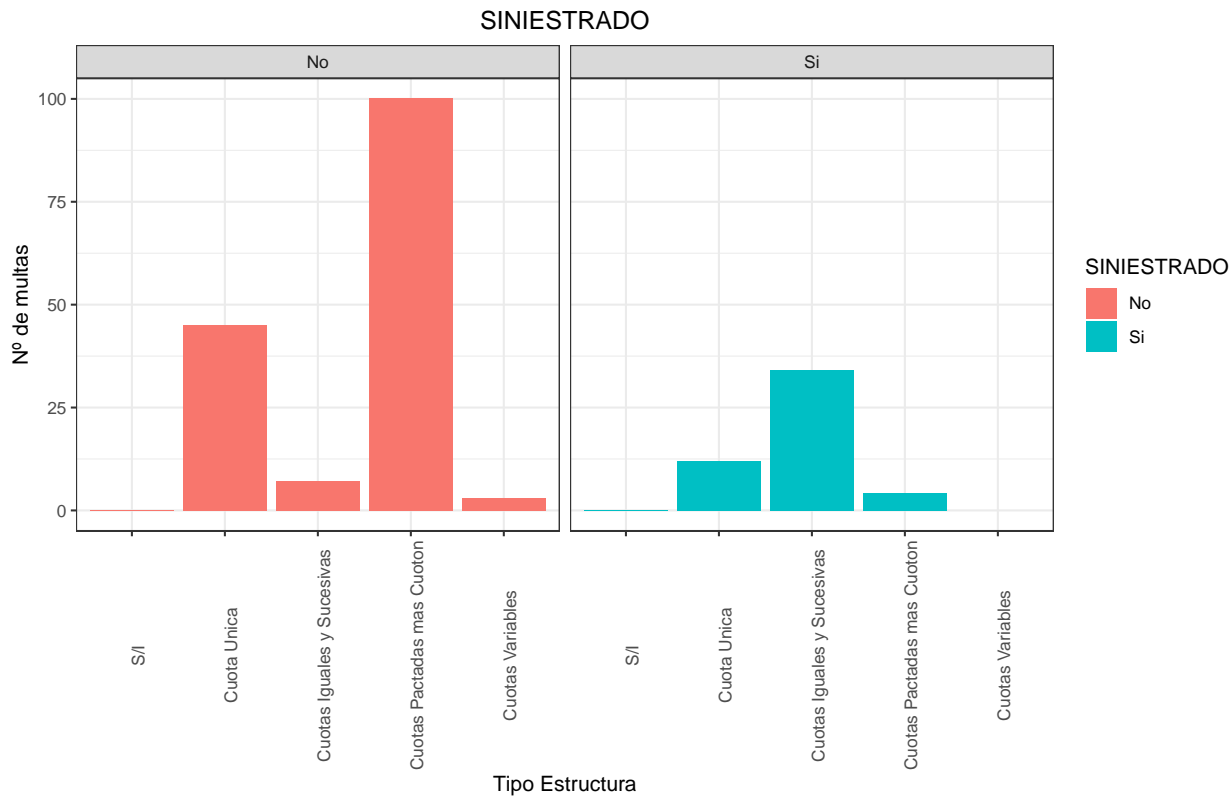


Figura 31

Por otra parte, clientes con altos valores de MONTO\_CERT\_PESOS y Monto\_Mora\_Pesos, son los que mayormente SINIESTRAN, pero a la misma vez, los clientes con altos valores de MONTo\_CERT\_PESOS y altos Monto\_Mora\_Pesos son NO SINIESTRADOS.

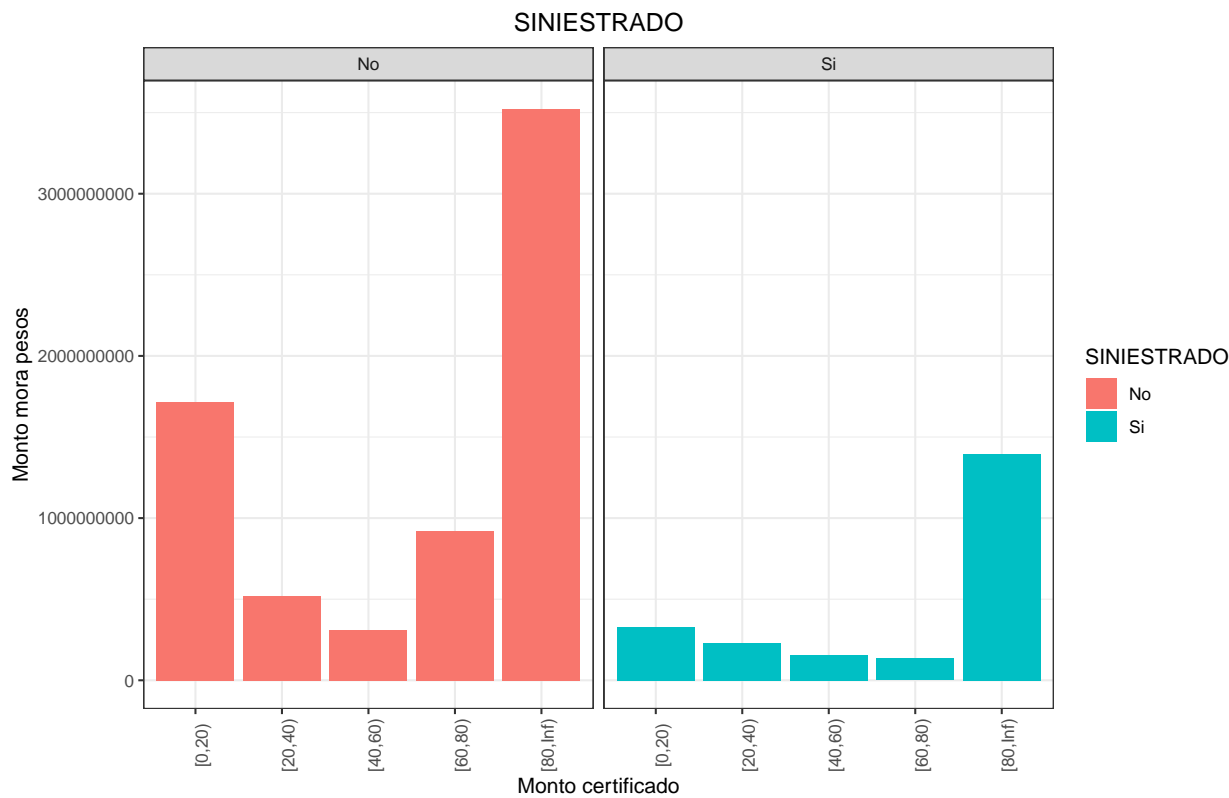


Figura 32

Finalmente, los clientes con N\_Multas (>80) y NUM\_CUOTAS (>20) son los que efectivamente SINIESTRAN. Esta proporción también se da en los clientes que NO SINIESTRAN. Por otra parte se aprecia que los clientes con número de cuotas entre las 5 -10 no existen registros de SINIESTROS.

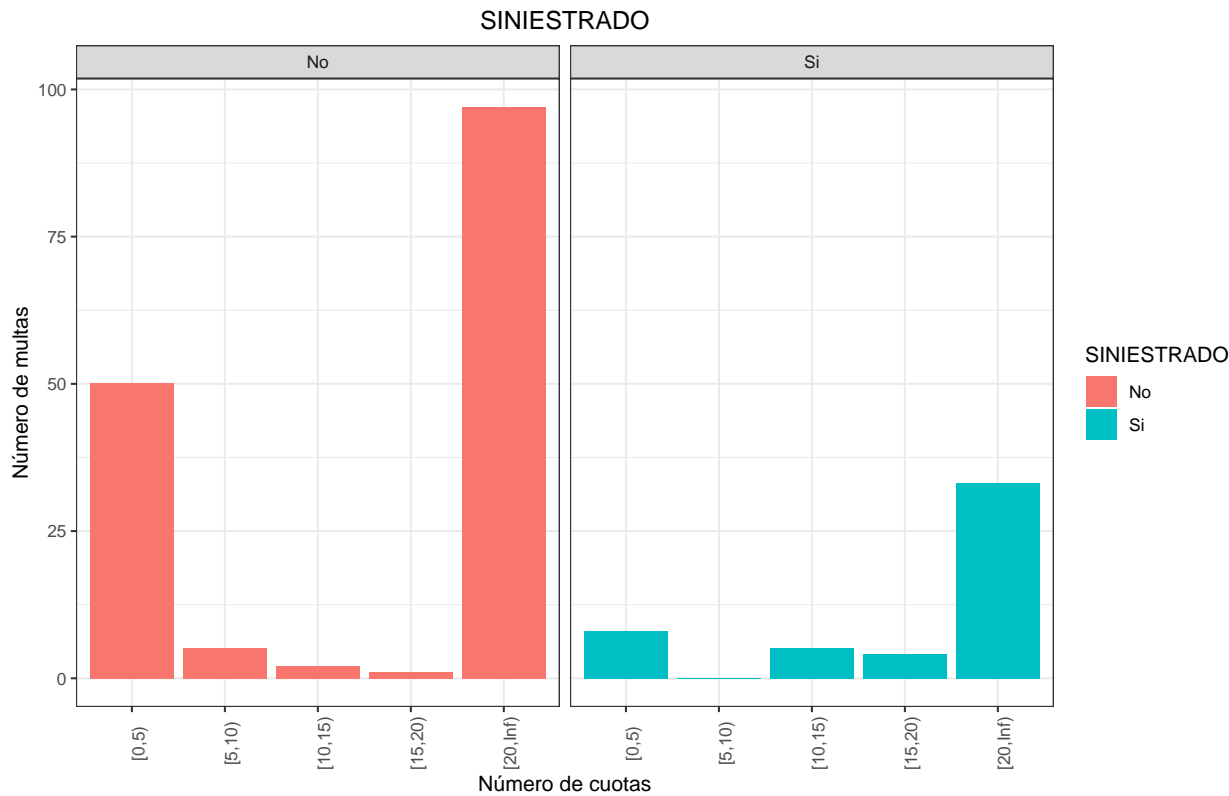


Figura 33

### 2.2.3. Análisis gráfico sobre SINIESTRADO2

Los siguientes análisis están basados en la segunda marca propuesta, la cual será representada como ‘SINIESTRADO2’ y que ya anteriormente fue descrita. La siguiente figura muestra que la proporción de la nueva marca propuesta, SINIESTRADO2, alcanza el 22.6% versus el 77.4% de los que NO SINIESTRAN. Esta proporción es más alta que en el anterior caso de SINIESTRADO (SINIESTRADO 13.7% y NO SINIESTRADO 86.3%). Cabe señalar que este no es un valor observado directamente, pero si intuitivo y consecuente desde la mirada comercial del problema, basado en el comportamiento de los clientes que pasan por el proceso de normalización.

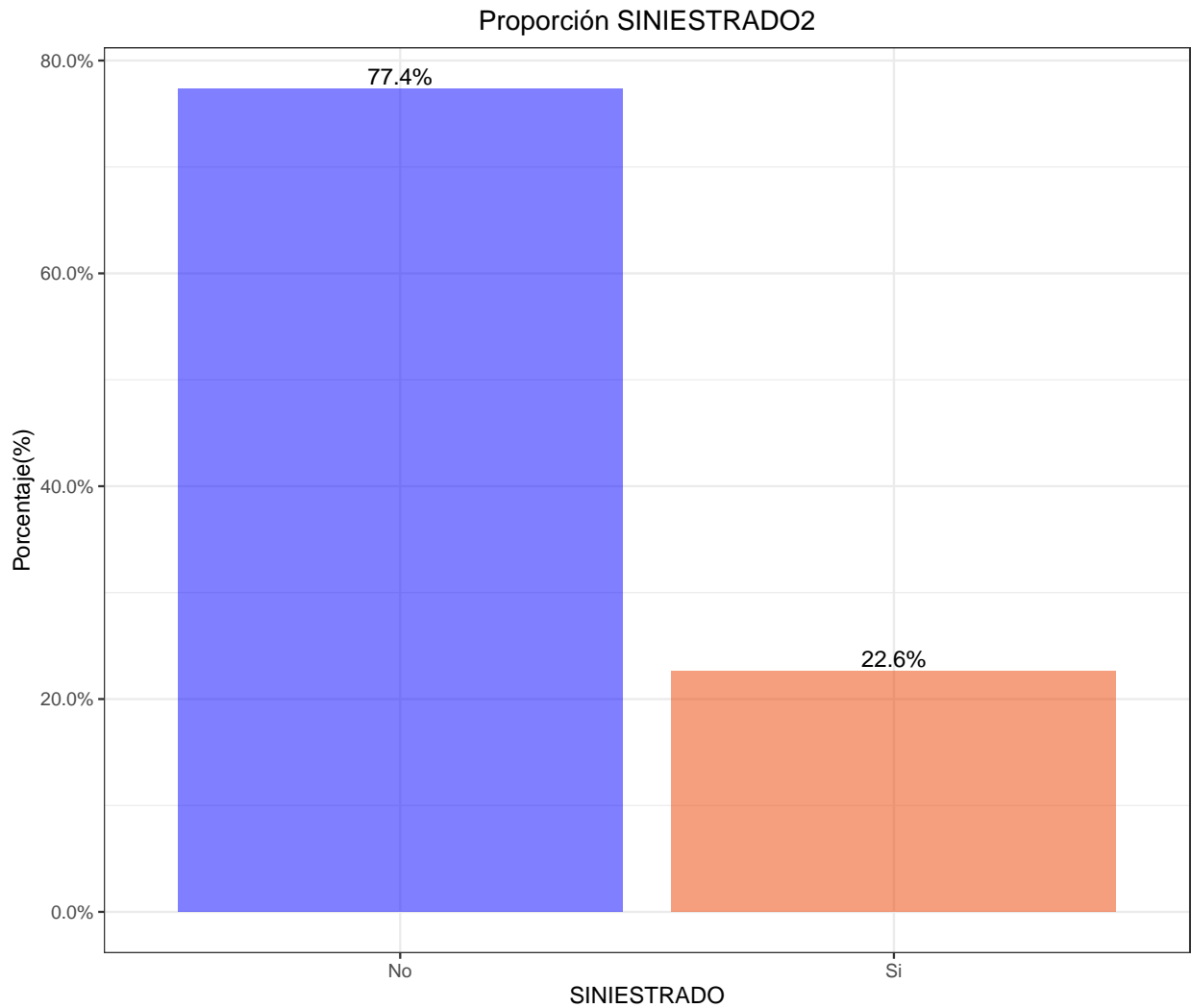
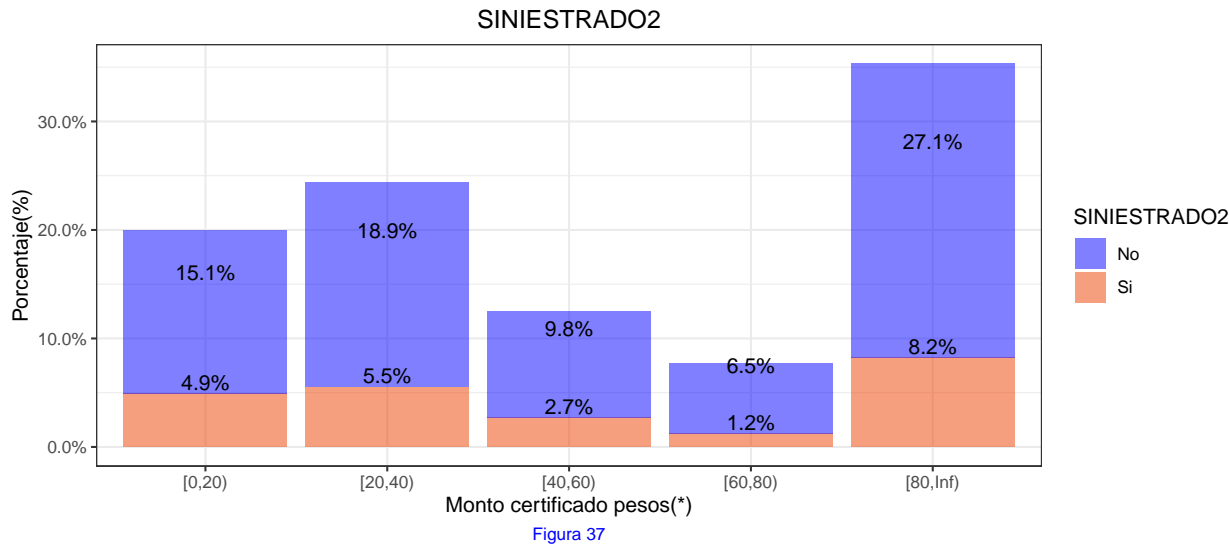
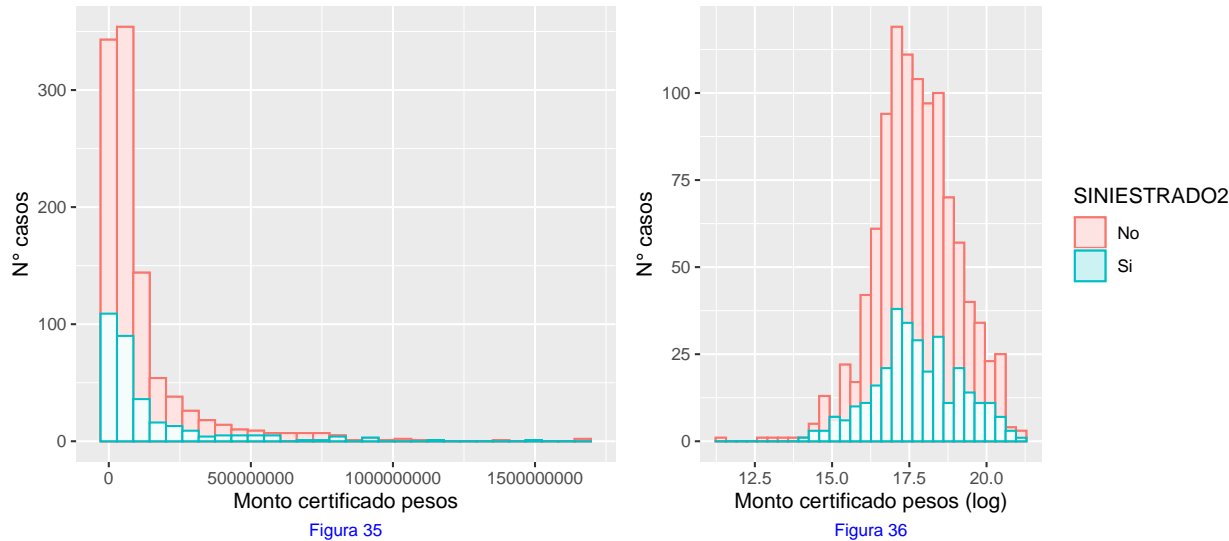


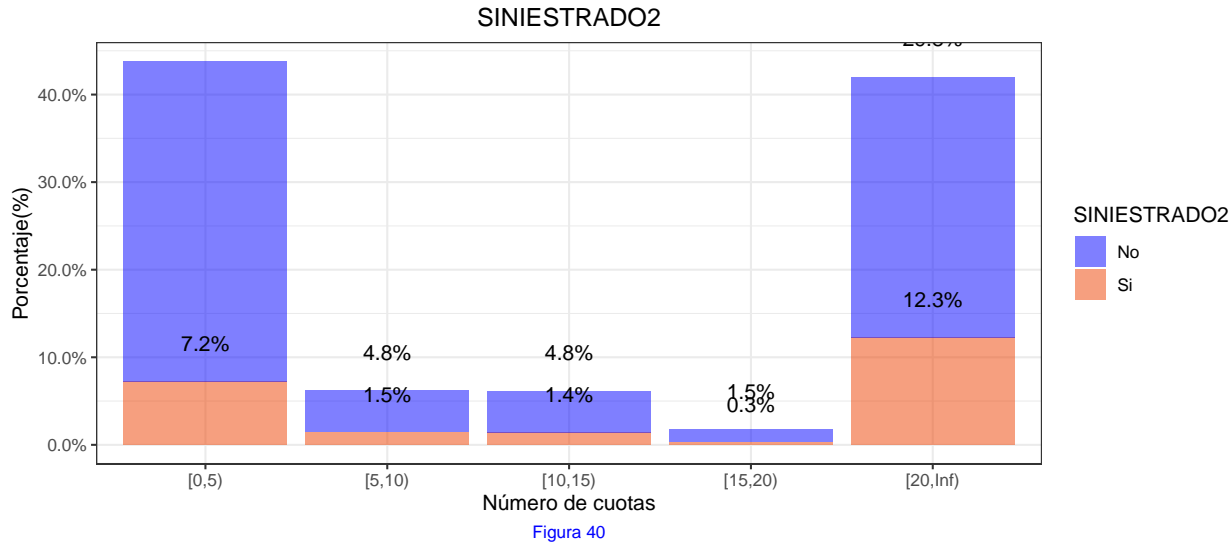
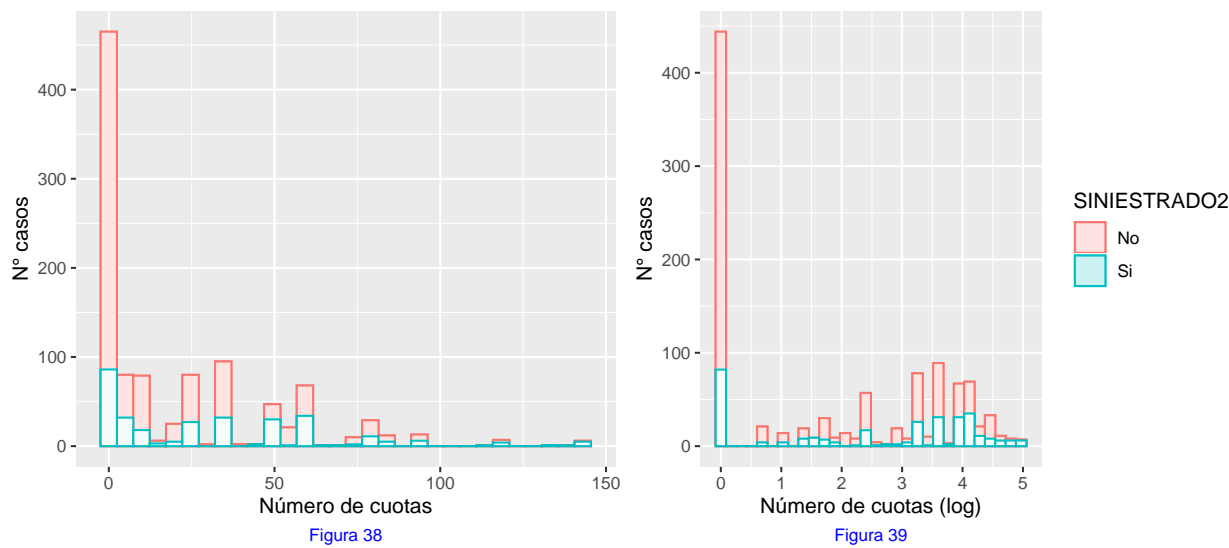
Figura 34

En la siguiente figura se muestra el mismo patrón de comportamiento para los clientes que SINIESTRAN / NO SINIESTRAN en función de la variable ya descrita anteriormente (MONTO\_CERT\_PESOS). El monto del certificado en pesos esta concentrado mayormente en el rango de los \$0 - \$500000000 y los clientes que tienen una mayor proporción de SINIESTROS se encuentra en el rango correspondiente a los montos más altos de certificados (mayores a \$800000000).

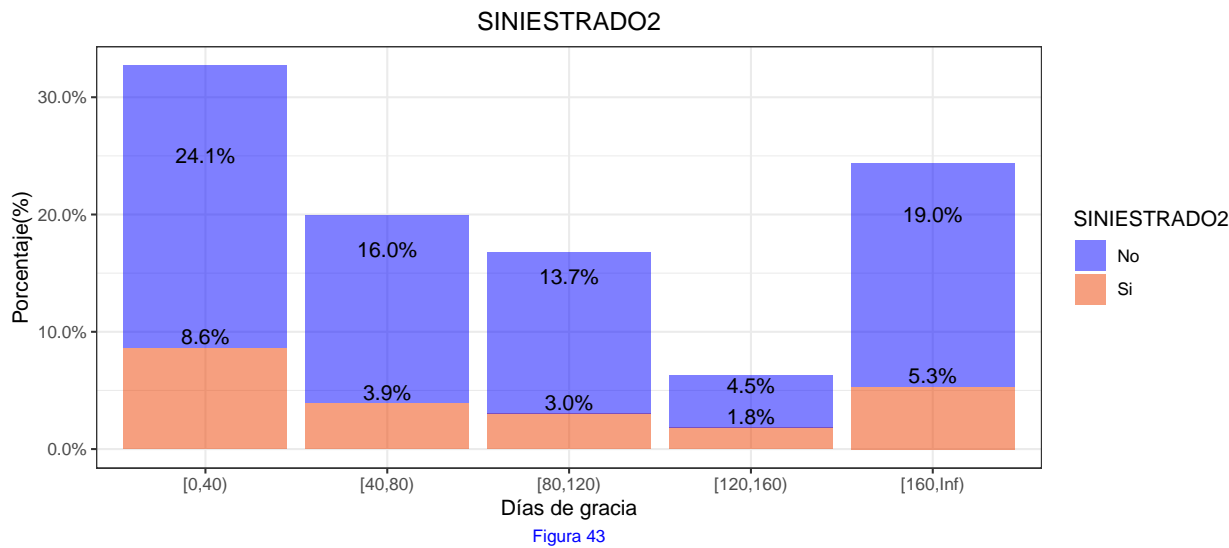
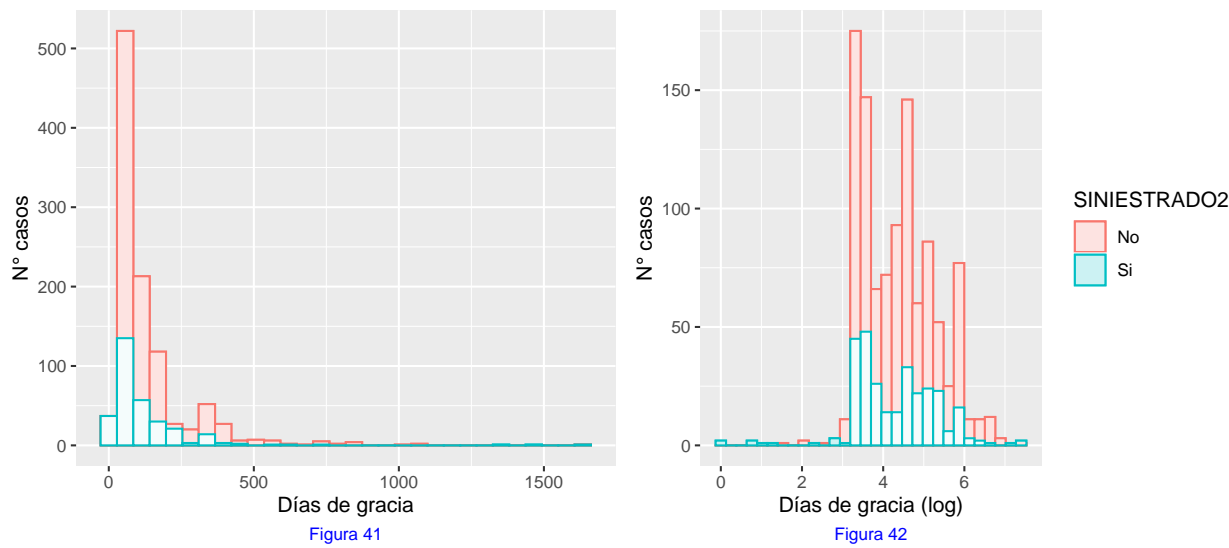


Nota: El gráfico anterior (Monto certificado en pesos), para representar de mejor forma la escala del eje x, el valor fue dividido por \$1.000.000.

El siguiente gráfico está relacionado al número de cuotas en el cual un cliente pacta su contrato. En este caso los clientes que SINIESTRAN en mayor proporción se encuentran entre las 20 o más número de cuotas pactadas, seguido por los clientes ubicados en los rangos de 0 -5 número de cuotas pactadas (7.2%). Al igual que en el caso anterior, las variaciones distribucionales en escala original y logarítmica no muestran ningún patrón de dispersión hacia valores que podrían tener un comportamiento atípico, sin embargo esta es una apreciación visual a priori, por lo que en Anexo II se puede revisar de mejor forma esta apreciación.

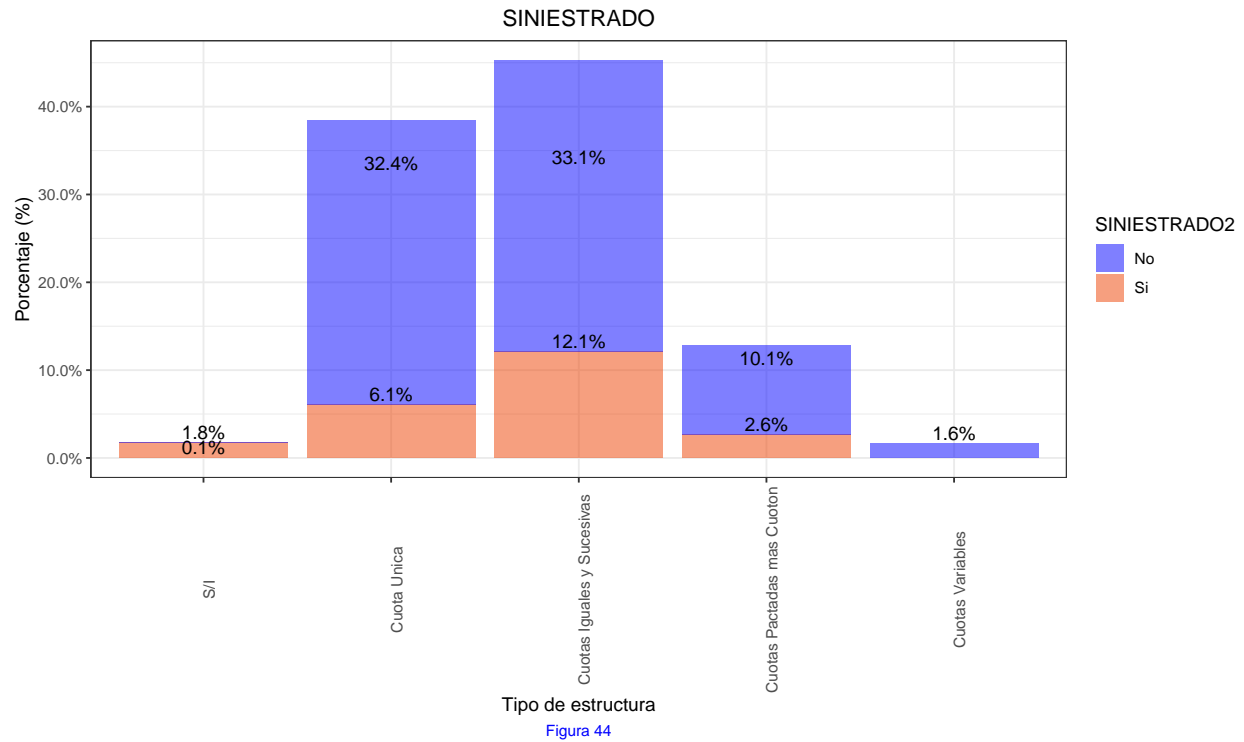


La variable DIAS\_GRACIA tiene el mismo comportamiento de las anteriores variables en términos distribucionales. En el rango de los 0 a 40 DIAS\_GRACIA es donde se concentra la mayor proporción de clientes SINIESTRADOS, seguido por el rango mayor a 160 días, y posteriormente en el rango de 40 - 80 días de gracia.

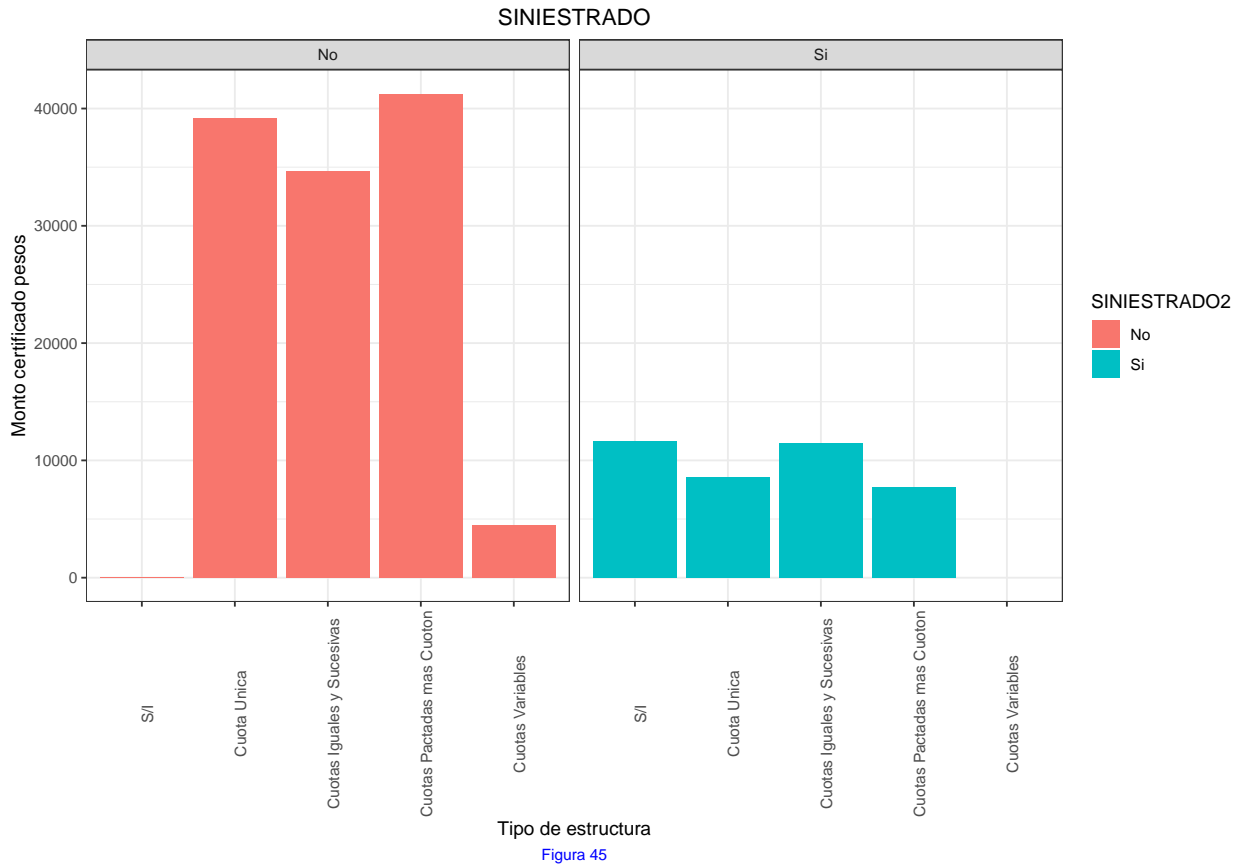




A continuación se presenta las variable categórica ‘TIPO\_ESTRUCTURA’ que se encuentra en ‘data\_cuadra’ pero relacionada en este caso a SINIESTRADO2. Se aprecia que el mayor pocentaje de los clientes SINIESTRADOS son aquellos que pactan un contrato con ‘Cuota Iguales y Sucesivas’ (12.1%), seguido de aquellos con contratos estructurados en ‘Cuota Única’ (6.1%) y de ‘Cuotas Pactadas Más Cuotón’ (2.6%). Aquellos clientes que pactan el contrato con ‘Cuotas Variables’, en este caso de SINIESTRADO2, no SINIESTRAN.



El gráfico siguiente muestra como, de acuerdo a la distribución del MONTO\_CERT\_PESOS, los clientes SINIESTRADOS presentan un comportamiento bastante similar entre los tipos de estructura de contratos. Es decir, el tipo de contrato no afecta de manera directa en que si un cliente SINIESTRA / NO SINIESTRA en el caso de la evaluación de SINIESTRADO2, a diferencia del primer caso SINIESTRADO.



Nota: El gráfico anterior (Monto de certificado), para representar de mejor forma la escala del eje x, el valor fue dividido por \$1.000.000.

Los siguientes gráficos están en función de las variables disponibles en ‘data\_equi’. El gráfico de abajo nos muestra la relación existente entre la variable Monto\_Mora\_Pesos y los SINIESTRADO/NO SINIESTRADO. Los clientes que presentan una mayor proporción de siniestralidad se encuentran en el rango de \$0 - \$20000000 de pesos con un 17.3% del total.

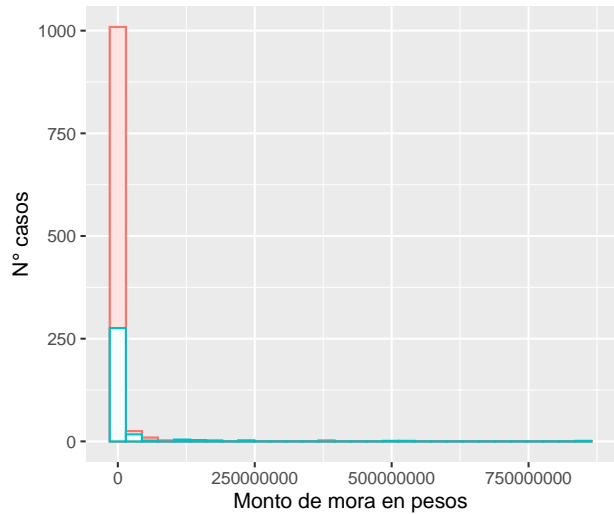


Figura 46



Figura 47

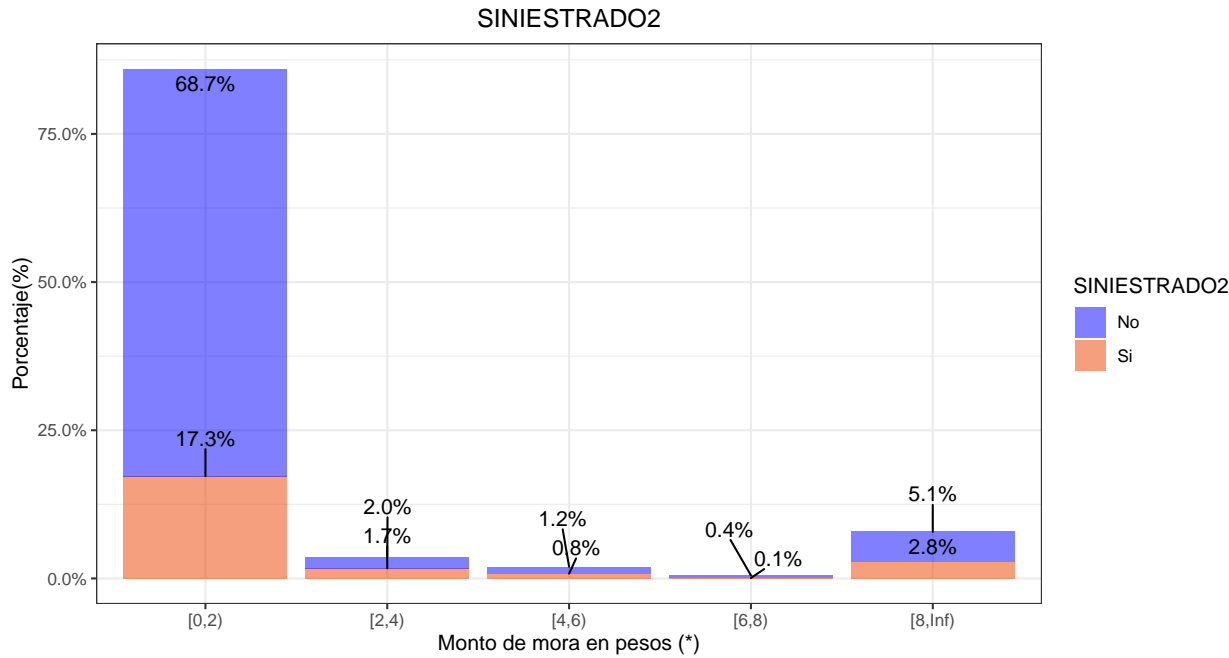


Figura 48

*Nota: El gráfico anterior (Monto de moras en pesos), para representar de mejor forma la escala del eje x, el valor fue dividido por \$1.000.000.*

El gráfico de N\_Moras también muestra homogeneidad entre los clientes SINIESTRADO/NO SINIESTRADO, pero el gráfico siguiente nos muestra que los clientes SINIESTRADOS con un mayor porcentaje se encuentran en el rango de las 0 - 1 mora (13.2%), mientras que el siguiente mayor número de clientes SINIESTRADO se concentra entre en el rango de 4 o más número de moras (4.6%). Tal como en en análisis previo, al parecer también existe presencia de valores atípicos, por los que este análisis particular se puede revisar en Anexo II.

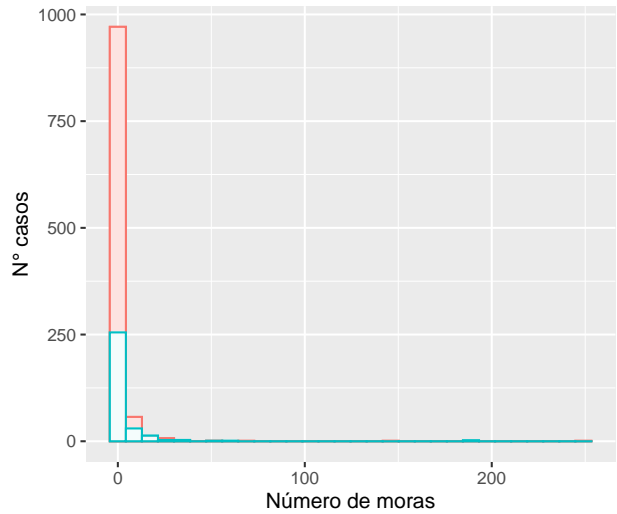


Figura 49

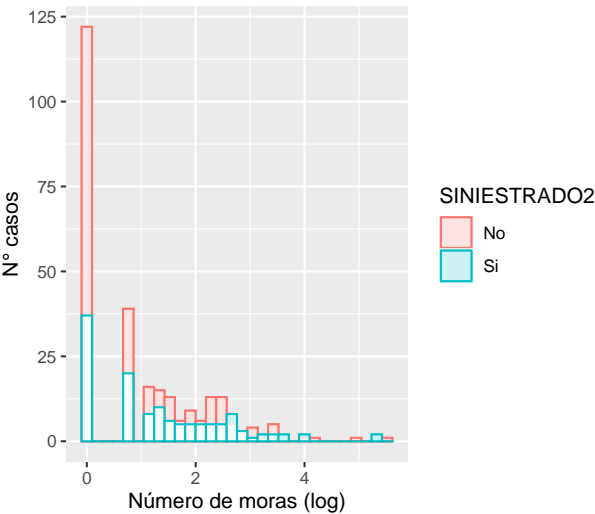


Figura 50

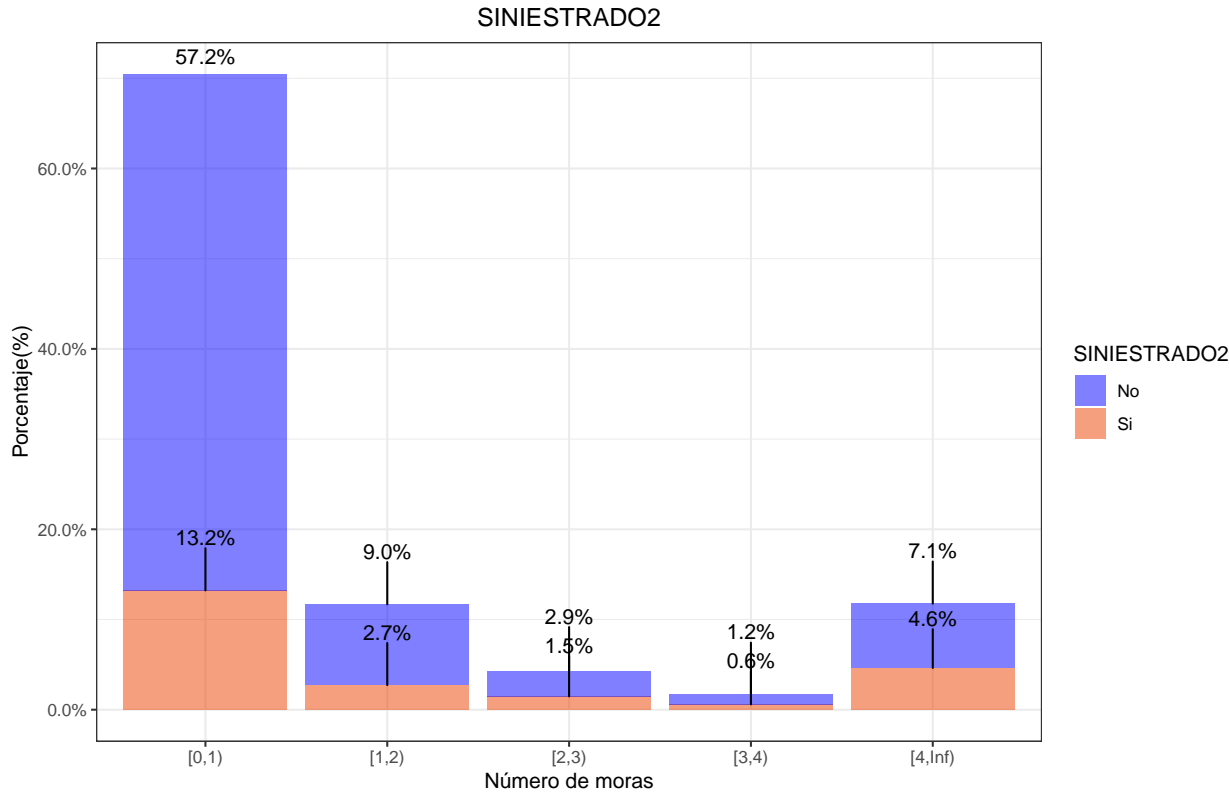


Figura 51

La variable N\_Multas no presenta mayor variabilidad en su distribución de SINIESTRADOS / NO SINIESTRADOS, visualmente tampoco se aprecian valores atípicos en ambas distribuciones (escala logarítmica y original), y ambas proporciones presentan el mismo patrón. En relación al porcentaje de SINIESTRADOS, la mayor cantidad de clientes se encuentra en el rango de 0 - 1 número de multas, concentrando casi la totalidad de la muestra para esta variable. Se puede apreciar que existe un valor cercano a las 70 número de multas por lo que este valor al parecer sería un outlier (ver Anexo II)

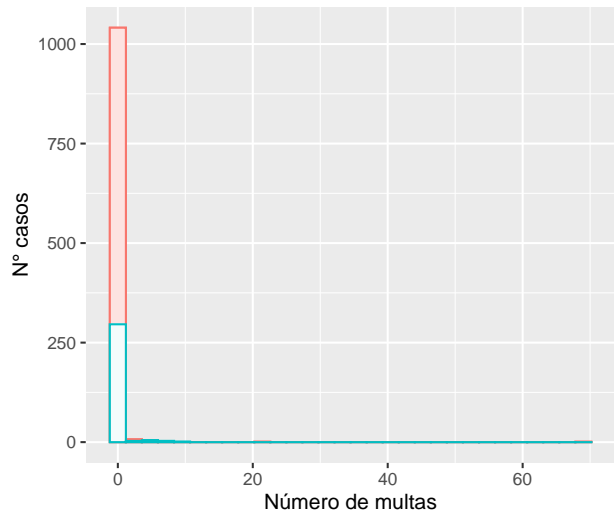


Figura 52

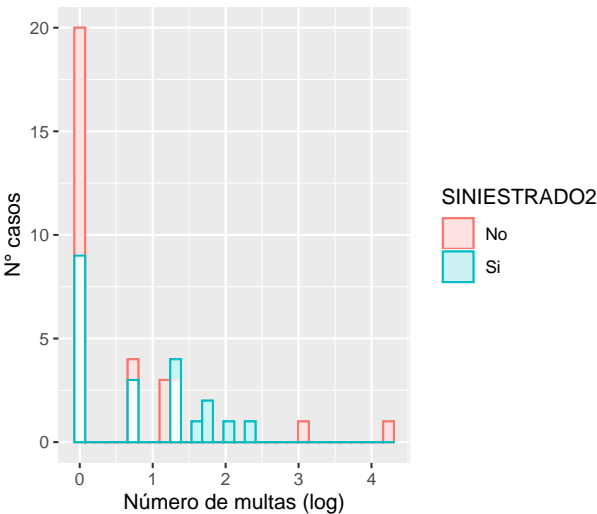


Figura 53

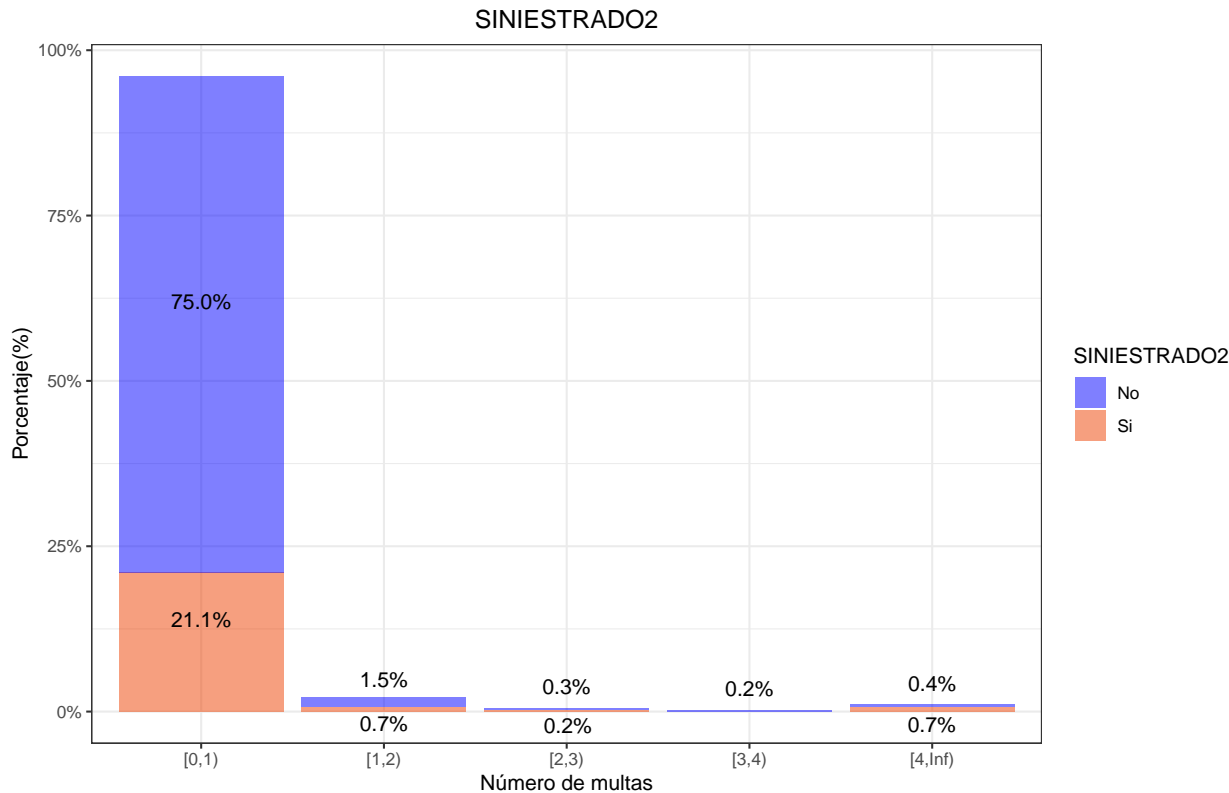


Figura 54

El gráfico del N\_Protestos también es homogéneo visualmente entre los clientes SINIESTRADO/NO SINIESTRADO. En los anteriores gráficos el mayor porcentaje de SINIESTRADO se concentra dentro del primer rango de distribución, en este caso dentro del 0 - 1 con un número de protestos (15.9%), seguido del rango de 3 o más número de protestos (5.4%)

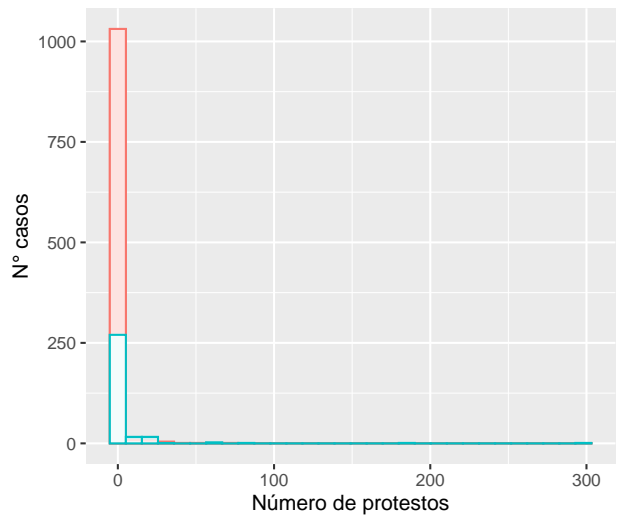


Figura 55

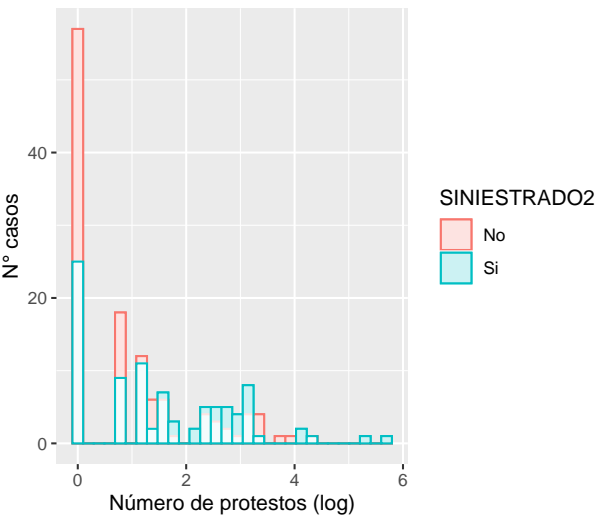


Figura 56

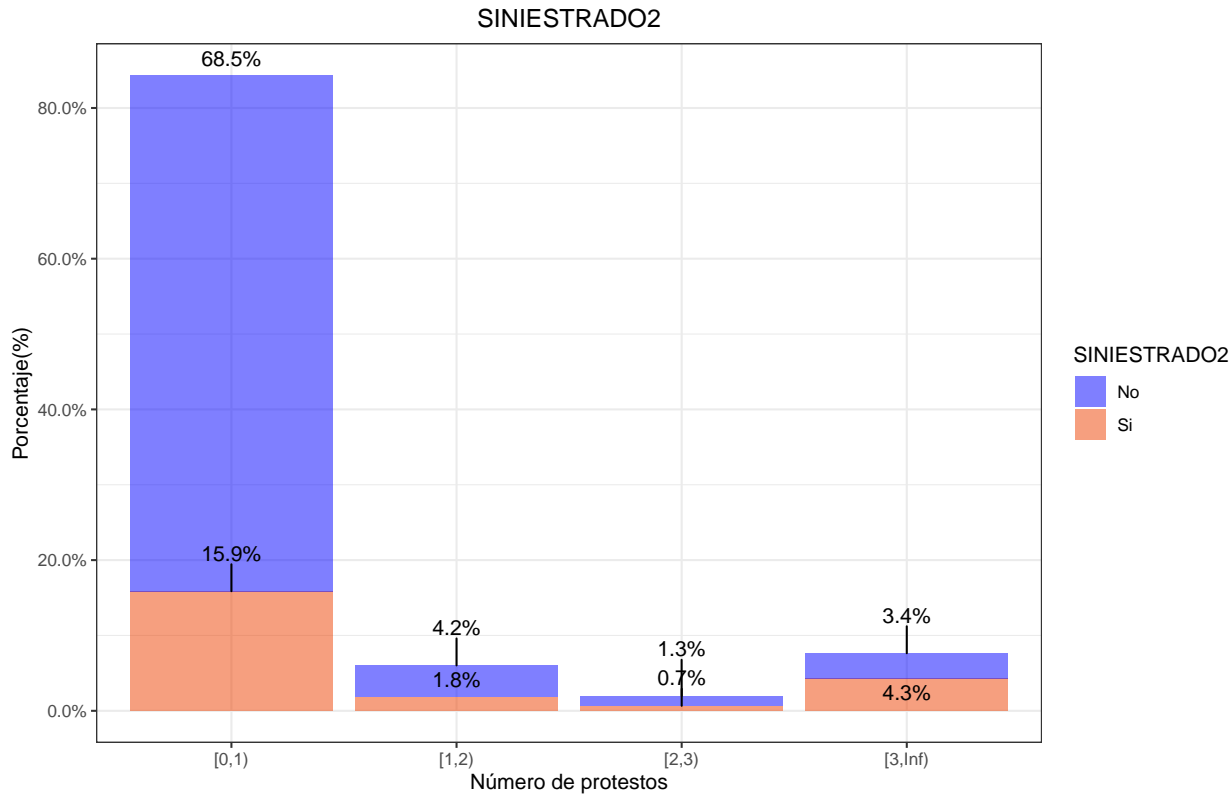


Figura 57

La relación TIPO\_ESTRUCTURA y N\_Multas refleja que los clientes con un número de multas alto y que tiene pactado su contrato en ‘Cuotas Iguales y Sucesivas’ son los que tienen mayor probabilidad de SINIESTRAR. En cambio, aquellos clientes con contrato tipo ‘Cuotas Pactadas más Cuotón’ con un número alto de multas, son los que tienen menos probabilidad de SINIESTRAR. Esto quiere decir que el número de multas no es un indicativo por si solo para evaluar si un cliente SINIESTRA / NO SINIESTRA.

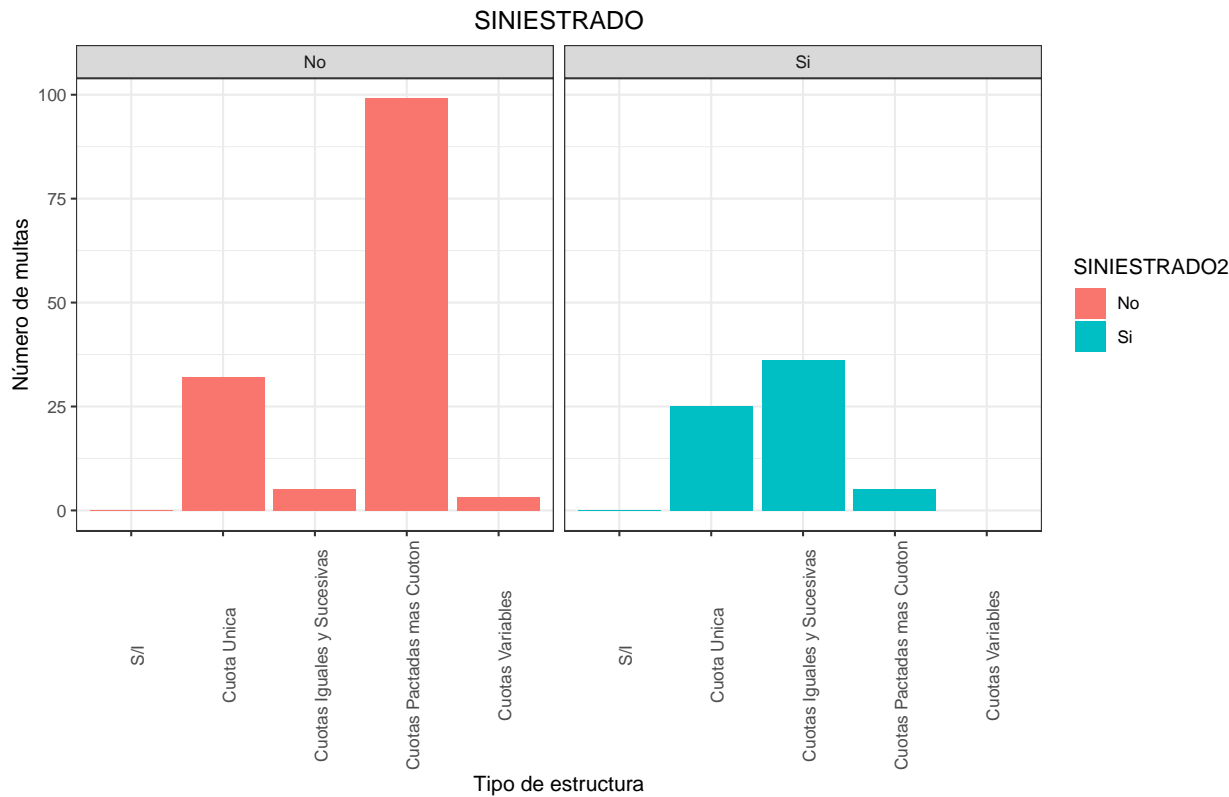


Figura 58

Por otra parte, clientes con un valor del MONTO\_CERT\_PESOS y Monto\_Mora\_Pesos altos, son los que mayormente SINIESTRAN. Pero, si vemos este mismo rango del MONTO\_CERT\_PESOS en los clientes que NO SINIESTRAN, son los que también tienen mayor un Monto\_Mora\_Pesos alto, es decir, ambas por si solas no son explicativas para determinar si un cliente tiene alta probabilidad de SINIESTRAR.

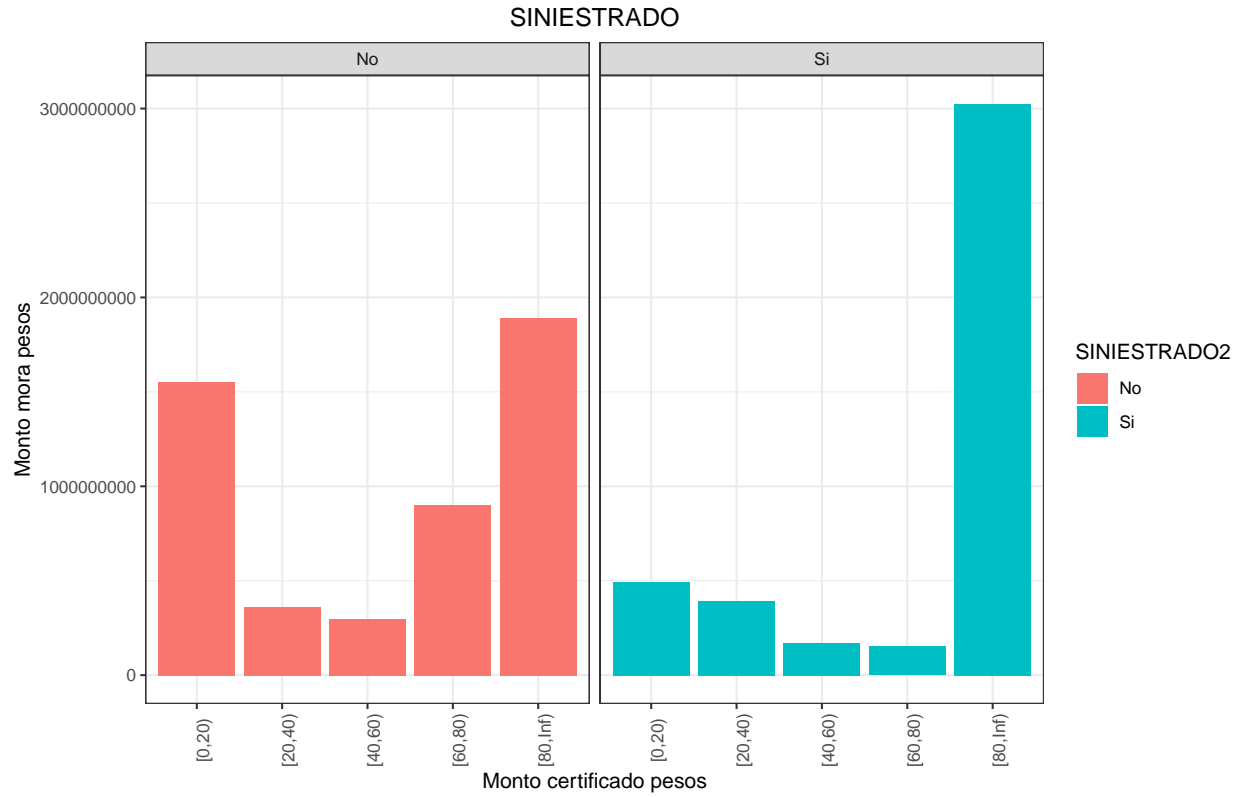


Figura 59



Finalmente, el NUM\_CUOTAS versus el N\_Multas nos indica que los clientes con mayor número de multas (>80) y mayor número de cuotas (>20) efectivamente son los que SINIESTRAN, pero esta misma proporción también se da en los clientes que NO SINIESTRAN, por tanto ocurre lo mismo que en los casos anteriores.

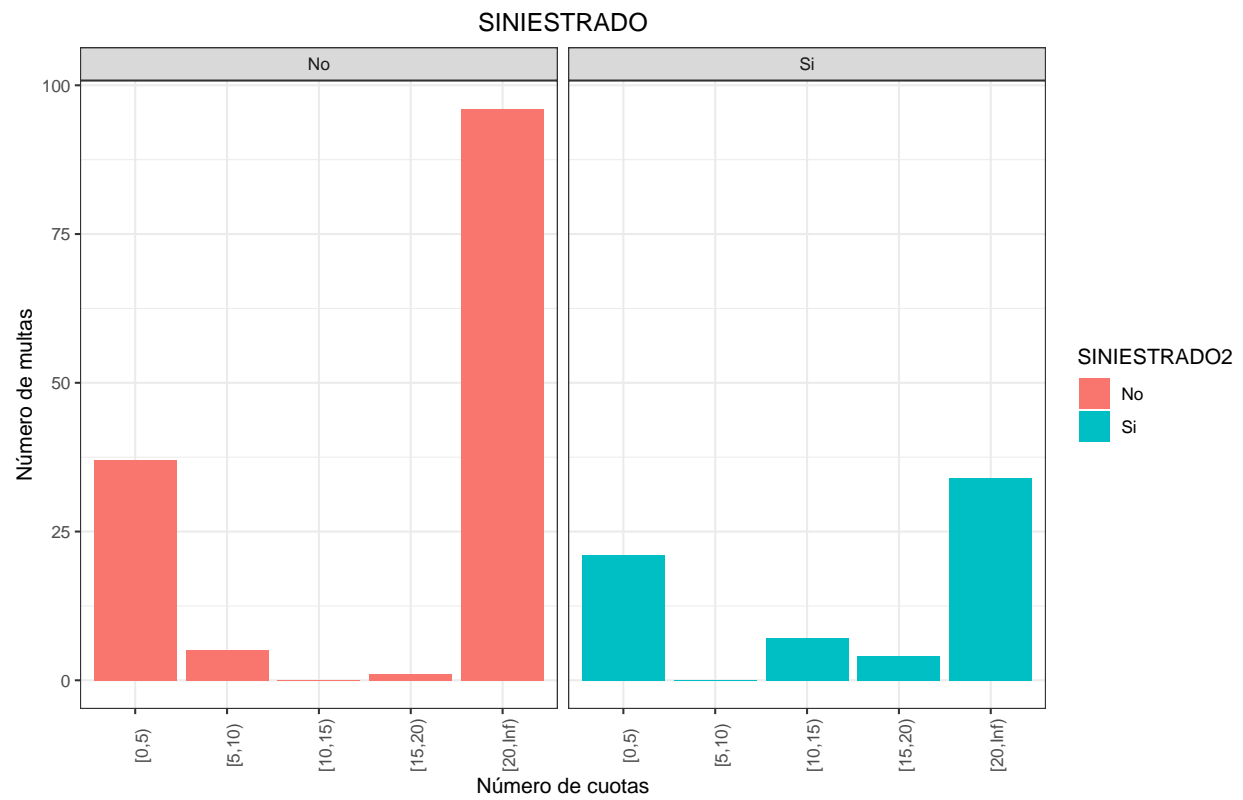


Figura 60

### 3. Modelo predictivo SINIESTRO de clientes

#### 3.1. Modelos de clasificación binaria

Una de las formas más comunes para obtener una probabilidad de ocurrencia ante una respuesta con sólo dos posibles resultados es a través de un modelo de clasificación binaria. Estos tipos de modelos responden a sólo dos posibles resultados: si/no, tiene/no tiene, hombre/mujer, etc.

Dentro del área de Machine Learning existe un amplio rango de técnicas para modelar problemas de clasificación binaria disponibles en la literatura, pero en este proyecto se evaluarán sólo 3, ya que para son los más utilizados y de mayor eficiencia computacional:

- I Modelo Lineal Generalizado (GLM)
- II Gradient Boosting Machine (GBM)
- III Distributed Random Forest (DRF)

Cabe recordar que estos modelos son propuestos en base a la información disponible que es considerada como “*data estructurada*”, por tanto, estos modelos corresponden a la clase de “modelos supervisados” o también llamados “aprendizaje supervisado”.

#### I. Modelo Lineal Generalizado (GLM)

Los Modelos Lineales Generalizados (GLM; McCullagh and Nelder (1989)) proporcionan una flexible generalización de los modelos lineales tradicionales los cuales asumen un error  $\epsilon$  de distribución  $\mathcal{N} \sim (0, 1)$ . Los GLM permiten modelar distribuciones de variables que pertenecen a la familia exponencial a través de una función de enlace que relaja el supuesto de normalidad en los errores de la distribución lineal clásica. Además, unifican una serie familias de distribuciones probabilísticas usualmente conocidas; Poisson, Binomial, Gamma, Normal, etc. Un caso especial de este tipos de modelos es la Regresión Logística, metodología usada generalmente para modelar variables con respuesta binaria, obteniéndose como resultado un rango de probabilidad estimada para cada observación. La expresión matemática para este tipo de modelos es la siguiente:

$$\begin{aligned} \mathbf{Y} &\sim \text{Bernoulli}(p_i) \\ \boldsymbol{\eta} &= \mathbf{X}\boldsymbol{\beta} \\ p_i &= h(\eta_i) \end{aligned}$$

donde  $h$  es una función invertible que mapea en el rango de probabilidades (0,1) dentro de  $\mathcal{R}$  y el vector  $\boldsymbol{\eta}$  es el predictor lineal. Para generar la linealidad  $h^{-1} = g$ , en donde  $g(\cdot)$  es la función de enlace, por lo tanto:

$$g(p) = \text{logit}(p) = \log\left(\frac{p}{1-p}\right) \quad (1)$$

y

$$h(\eta) = g^{-1}(\eta) = \frac{1}{1 + e^{-\eta}} \quad (2)$$

donde la ecuación (1) es llamada la función *logit* y la ecuación (2) es la inversa de la función *logit*

La estimación de parámetros se hace mediante el método de máxima verosimilitud (log-likelihood):

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (3)$$

## II. Gradient Boosting Machine (GBM)

El algoritmo GMB es un clasificador que construye un modelo con predicción “fuerte” a partir de muchos modelos de predicciones “débiles” (Hastie *et al.*, 2005). La construcción se va construyendo en pasos iterativos y en forma secuencial combinando (*ensembling*) modelos sencillos de clasificación, en donde cada nuevo modelo que se incorpora al conjunto intenta disminuir los errores de las estimaciones anteriores, así de las múltiples combinaciones, el GBM genera relaciones no lineales entre la variable respuesta y los predictores. Gradient Boosting Machine (GBM) utiliza el descenso de la gradiente para optimizar la función de minimización (o maximización) durante el ajuste del modelo. Como observación final cabe señalar que el GBM utiliza como predicciones “debiles” a los modelos basados en árboles de decisión.

El algoritmo de clasificación es el siguiente:

---

**Algorithm 1** Gradient boosting machine

---

- 1: Inicializar la función de partida  $f_0$  con alguna constante y para  $t = 1$  hasta  $N$
- 2: Calcular la gradiente negativa de  $g_t(x)$
- 3: Ajustar el modelo base  $h(x, \theta_t)$
- 4: Encontrar la gradiente descendiente  $p_t$

$$p_t = \operatorname{argmin} \sum \psi[y_i, f_{t-1}(x_i) + p h(x_i, \theta_t)]$$

- 5: Actualizar la función estimada en el paso 1 mediante la gradiente y en el modelo base

$$f \leftarrow f_{t-1} + p_t h(x, \theta_t)$$

- 6: Finalizar
- 

El modelo base  $h(x, \theta)$  puede tener estructura para un modelo de regresión o clasificación. La función de pérdida  $\psi(y, f)$  se elige generalmente por las características de la variable de respuesta. En el caso de clasificación binaria se utiliza una función de pérdida binomial. Para evitar un sobre-ajuste, se pondera cada iteración y se utiliza un submuestreo aleatorio en los datos donde el modelo base  $h(x, \theta)$  es entrenado.

## III. Distributed Random Forest (DRF)

El modelo de clasificación Distributed Random Forest (DRF) es un algoritmo de aprendizaje conjunto que utiliza árboles de decision para la clasificación de variables binarias. Su estructura es muy sencilla ya que, en simples palabras, combina múltiples árboles de decision para formar un clasificador final. A través de un conjunto de múltiples clasificadores en variables no correlacionadas, las diferencias en las estimaciones de los árboles (sesgo de estimación) ocurrirán debido a las variaciones propias entre los árboles. El algoritmo de RF es sencillo y sigue los siguientes pasos (Hastie *et al.*, 2002):

---

**Algorithm 2** Random Forest

---

- 1: Obtener mediante un Bootstrap muestras aleatorias *iid* del conjunto de datos de entrenamiento
  - 2: Generar un árbol de desición desde los datos muestreados. En cada nodo se muestrean aleatoriamente los predictores y se elige la mejor división entre los escogidos
  - 3: Predecir nuevos datos mediante la agregación de la predicción en la división
- 

La aleatoriedad en el re-muestreo de cada árbol garantiza que la varianza sea baja y que el sesgo (bias) no aumente en cada estimación. El Random Forest también permite extraer del modelo elegido a las variables con mayor importancia en función de explicar la variable respuesta.

### 3.2. Evaluación de los modelos

Para evaluar la performance de cada modelo se proponen las métricas comunmente utilizadas en este tipo de problemas, todas en función de las predicciones de cada uno de ellos y expresadas en la matriz de confusión.

Primero detallamos la notación en las predicciones:

- TP = True positive (Verdadero positivo)  $\Rightarrow$  Observación es clasificada como positiva cuando esta sí es positiva
- FP = False positive (Falso positivo)  $\Rightarrow$  Observación es clasificada como positiva cuando esta es negativa
- TN = True negative (Verdadero negativo)  $\Rightarrow$  Observación clasificada como negativa cuando esta sí es negativa
- FN = False negative (Falso negativo)  $\Rightarrow$  Observación clasificada como negativa cuando esta es positiva

De lo anterior podemos calcular métricas basadas en las predicciones anteriores:

Table 4: Métricas de evaluación para los modelos propuestos

Métrica	Cálculo
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$
Precision	$\frac{TP}{TP+FP}$
Sensibilidad ( <i>Recall</i> )	$\frac{TP}{TP+FN}$
Especificidad	$\frac{TN}{TN+FP}$

### 3.3. Curva ROC y AUC

La curva ROC permite visualmente analizar el rendimiento del modelo de clasificación. La curva ROC gráfica la tasa de TP (Sensibilidad) versus la tasa de FP (1 - Especificidad) para diferentes puntos de corte (James *et al.*, 2013).

El Área bajo la Curva (AUC) es una forma de resumir el rendimiento de un modelo con un solo valor. El valor es el área bajo la curva ROC y es una relación entre 0 y 1 donde un valor de 1 es un clasificador perfecto, mientras que un valor cercano a 0.5 es un modelo con una baja clasificación ya que esta clasificación sería casi de forma aleatoria (James *et al.*, 2013).

### 3.4. Plataforma de modelado

El modelo se construirá en el software estadístico R y se utilizarán principalmente las siguientes librerías:

- `sparklyr`
- `h2o`

La librería `sparklyr` genera una conexión con `Spark` y así realizar análisis estadísticos ante una gran cantidad de datos, permitiendo que el costo computacional de estimación sea considerablemente bajo y no colapsando la unidad de trabajo.

La librería `h2o` permite utilizar las técnicas de machine learning para la construcción del modelo de clasificación y su principal característica es que ejecuta los algoritmos en forma paralela utilizando todos los núcleos del computador (para optimizar los tiempos de ejecución).

## 4. RESULTADOS

### 4.1. SINIESTRADO

A continuación se presentan los resultados para la marca ‘SINIESTRADO’ comparando los 3 tipos de modelos propuestos. En primera instancia se utilizó la base unificada ‘full\_join’ (del esquema presentado en la primera figura) la cual integra información de ‘data\_cuadra’ y ‘data\_equi’. El conjunto de datos ‘full\_join’ fue dividido en 3: 70% para entrenar el modelo (data\_training), 15% para testear las predicciones (data\_test) y un 15% para validación (en términos de optimización de hiperparámetros y validación cruzada).

Modelo logístico (GLM)	
Predictores	Variable respuesta
MONTO_CERTIFICADO	SINIESTRADO
NUM_CUOTAS	
diff_months	
N_Protestos	
TIPO_ESTRUCTURA	
DIAS_GRACIA	
Monto_Mora_Pesos	
N_cert	
N_Multas	
N_Infracciones_Previsionales	

Statistics	data_training	data_test	data_val
MSE:	0.110	0.077	0.118
RMSE:	0.332	0.277	0.344
LogLoss:	0.362	0.275	0.387
Mean Per-Class Error	0.320	0.290	0.340
AUC:	0.753	0.810	0.687
pr_auc:	0.328	0.267	0.232
Gini:	0.506	0.620	0.374
R^2:	0.100	0.083	0.063

Gradient Boosting Machine (GBM)	
Predictores	Variable respuesta
MONTO_CERTIFICADO	SINIESTRADO
NUM_CUOTAS	
diff_months	
N_Protestos	
TIPO_ESTRUCTURA	
DIAS_GRACIA	
Monto_Mora_Pesos	
N_cert	
N_Multas	
N_Infracciones_Previsionales	

Statistics	data_training	data_test	data_val
MSE:	0.120	0.125	0.086
RMSE:	0.346	0.354	0.293
LogLoss:	0.399	0.416	0.319
Mean Per-Class Error	0.131	0.347	0.370
AUC:	0.940	0.713	0.645
pr_auc:	0.708	0.266	0.119
Gini:	0.880	0.427	0.291
R^2:	0.024	0.007	-0.028

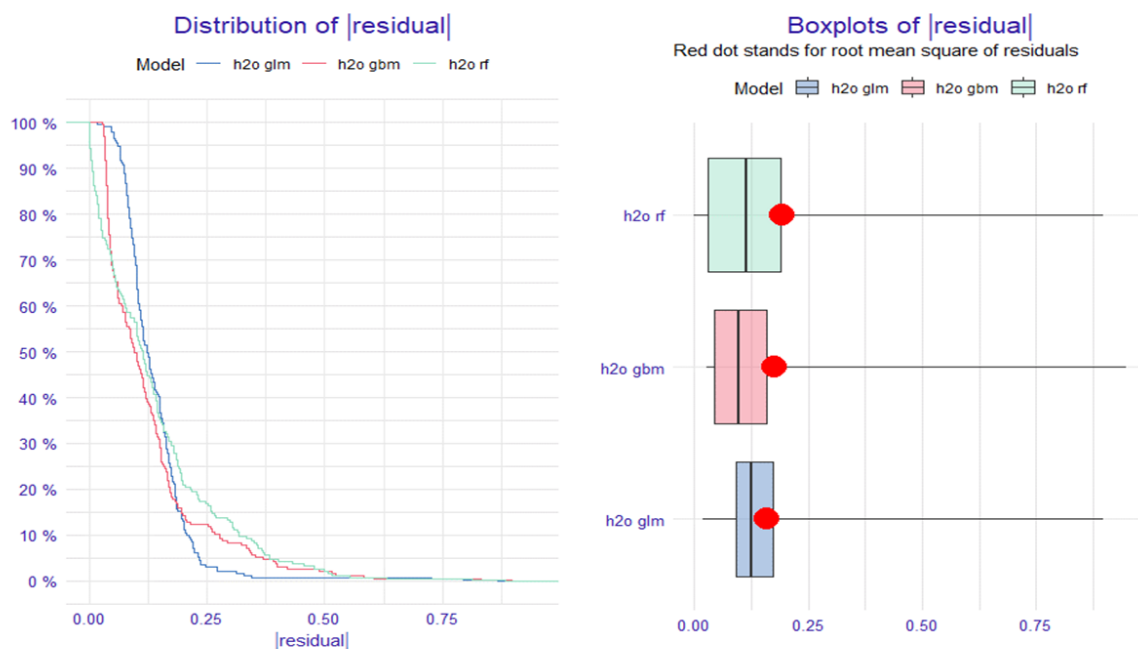
Distributed Random Forest (DRF)	
Predictores	Variable respuesta
MONTO_CERTIFICADO	SINIESTRADO
NUM_CUOTAS	
diff_months	
N_Protestos	
TIPO_ESTRUCTURA	
DIAS_GRACIA	
Monto_Mora_Pesos	
N_cert	
N_Multas	
N_Infracciones_Previsionales	

Statistics	data_training	data_test	data_val
MSE:	0.108	0.111	0.081
RMSE:	0.328	0.333	0.285
LogLoss:	0.351	0.358	0.283
Mean Per-Class Error	0.308	0.279	0.340
AUC:	0.766	0.768	0.753
pr_auc:	0.345	0.372	0.196
Gini:	0.532	0.536	0.506
R^2:	0.121	0.122	0.030

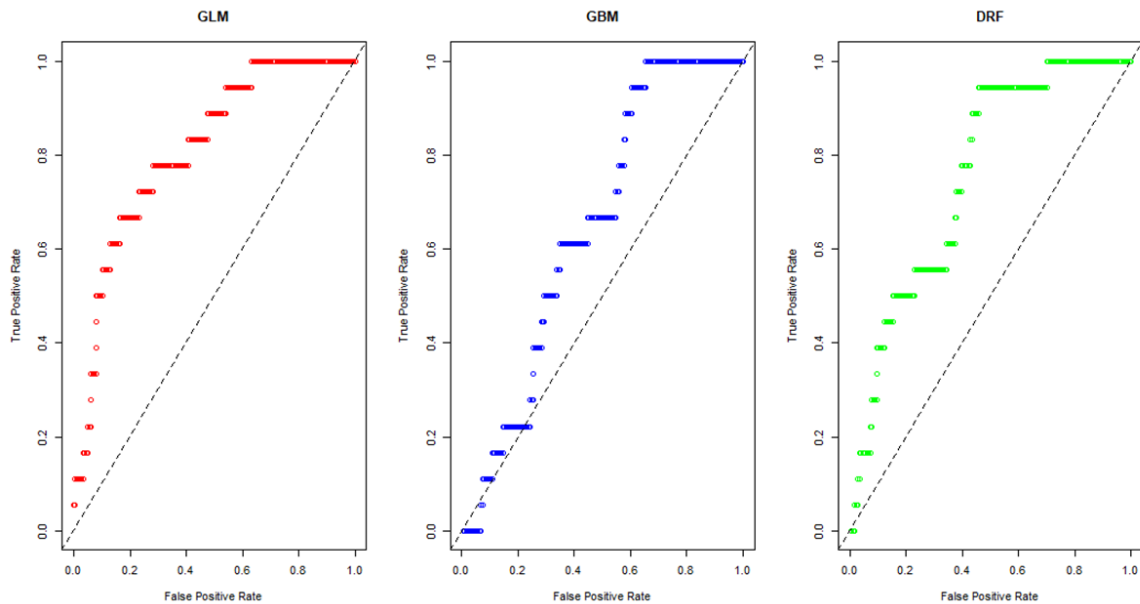
Como se puede apreciar en las tablas anteriores, el Distributed Random Forest (DRF) es el método que no generaliza el ajuste (aprendizaje) sobre los datos observados en comparación con las otras dos metodologías propuestas (GLM y GBM), ya que en terminos del AUC está por encima del valor 0.75 (que es bueno para estos casos) y además es estable entre los tres conjuntos de datos (data\_training, data\_val y data\_test).

También se empleó un test de diagnóstico para evaluar los modelos predictivos relacionado a los residuales de las predicciones. En la figura a continuación se que aprecia que la distribución de los residuales en los cuantiles es similar en los 3 modelos, teniendo una mediana menor en el modelo GMB seguido por el modelo DRF, para tener una menor performance el modelo GLM. Observando los diagramas de caja, también se puede ver que el modelo GBM tiene el valor residual absoluto medio más bajo, por lo tanto, aunque el modelo GBM tuvo el puntaje AUC más bajo, en realidad funciona mejor cuando se considera la mediana de los residuos absolutos.

Por otra parte, en la distribución de residuales del modelo GLM, se aprecia que este tiene una menor distribución hacia el lado derecho, indicando que, aunque su performance no es mejor que los modelos anteriores, la cantidad de residuos en la predicción es menor a los otros dos.



La siguiente figura muestra las curvas ROC sobre el conjunto el 'data\_test' para cada modelo propuesto. El que presenta una mejor curva ROC es el modelo GLM ya que el delta entre el área de la curva y la línea trazada en medio del gráfico es mayor, por tanto la predicción no parece ser de forma aleatoria (valor de AUC = 0.5). Esto va en directa concordancia con el gráfico de residuales presentado anteriormente que indica que la distribución de los residuos en las predicciones sobre el 'data\_test' es mejor en el GLM. El segundo modelo con mejor performance en esta métrica es el DRF que también tiene un delta mayor al GBM en relación al área sobre el corte del gráfico 50/50. Cabe señalar que los diagnósticos se hacen sólo sobre el 'data\_test' ya que es en este conjunto de datos en donde se predice el futuro comportamiento del cliente, el conjunto de 'data\_val' sólo nos permite encontrar, dentro del espacio paramétrico, el mejor conjunto de valores para los parámetros que son usados con fines de predictivos. Por lo anterior es que mismos valores de parámetros encontrados son utilizados para predecir sobre el 'data\_test'.



Finalmente se presenta la matriz de confusión para el conjunto de test (`data_test`), la cual nos permite analizar las predicciones correctas sobre este junto a sus respectivas métricas de evaluación. Como se puede apreciar, el modelo que presenta una mejor performance de predicción es el GLM, esto resumido en la precisión de la predicción (0.882) y el recall (0.500). El modelo GLM presenta muy buena performance cuando se trata de predecir sobre el ‘`data_test`’, con un AUC del 0.81, seguido del DRF con un 0.768 y al GBM con un 0.713, lo que se ve reflejado en el valor del accuracy. Los valores en las filas son las observaciones reales y los valores en las columnas son los predichos por el modelo, además, se recuerda que todas las métricas de evaluación están calculadas sobre los valores del ‘`data_test`’.

GLM				GBM				DRF			
		Predicciones				Predicciones				Predicciones	
		No	Si			No	Si			No	Si
Observaciones	No	163	14	Observaciones	No	115	62	Observaciones	No	155	22
	Si	9	9		Si	7	11		Si	10	8

Modelos	Accuracy	Precision	Recall	Especificidad
GLM	0.882	0.391	0.500	0.921
GBM	0.646	0.151	0.611	0.650
DRF	0.836	0.267	0.444	0.876

## 4.2. SINIESTRADO2

Los resultados a continuación están relacionados con la segunda marca propuesta para modelar, SINIESTRADO2, la cual está basada en los criterios presentados anteriormente. Se realizaron las mismas proporciones para la división del conjunto de datos, esto es: 70% datos entrenamiento (data\_training), 15% datos para test (data\_test) y 15% para validación (data\_val). Los resultados del modelado se presentan en las siguientes tablas, considerando las métricas de evaluación para cada una de las metodologías propuestas, resaltando en color amarillo el área bajo la curva (AUC) en las predicciones de cada uno de los modelos.

Modelo logístico (GLM)					
Predictores	Variable respuesta	Statistics	data_training	data_test	data_val
MONTO_CERTIFICADO	SINIESTRADO	MSE:	0.153	0.147	0.146
NUM_CUOTAS		RMSE:	0.392	0.384	0.381
diff_months		LogLoss:	0.478	0.462	0.456
N_Protestos		Mean Per-Class Error	0.330	0.302	0.328
TIPO_ESTRUCTURA		AUC:	0.721	0.706	0.703
DIAS_GRACIA		pr_auc:	0.480	0.486	0.429
Monto_Mora_Pesos		Gini:	0.443	0.412	0.407
		R^2:	0.131	0.166	0.124

Gradient Boosting Machine (GBM)					
Predictores	Variable respuesta	Statistics	data_training	data_test	data_val
MONTO_CERTIFICADO	SINIESTRADO	MSE:	0.171	0.174	0.164
NUM_CUOTAS		RMSE:	0.413	0.417	0.405
diff_months		LogLoss:	0.522	0.530	0.509
N_Protestos		Mean Per-Class Error	0.200	0.348	0.364
TIPO_ESTRUCTURA		AUC:	0.912	0.690	0.672
DIAS_GRACIA		pr_auc:	0.687	0.357	0.328
Monto_Mora_Pesos		Gini:	0.823	0.380	0.344
		R^2:	0.033	0.018	0.012

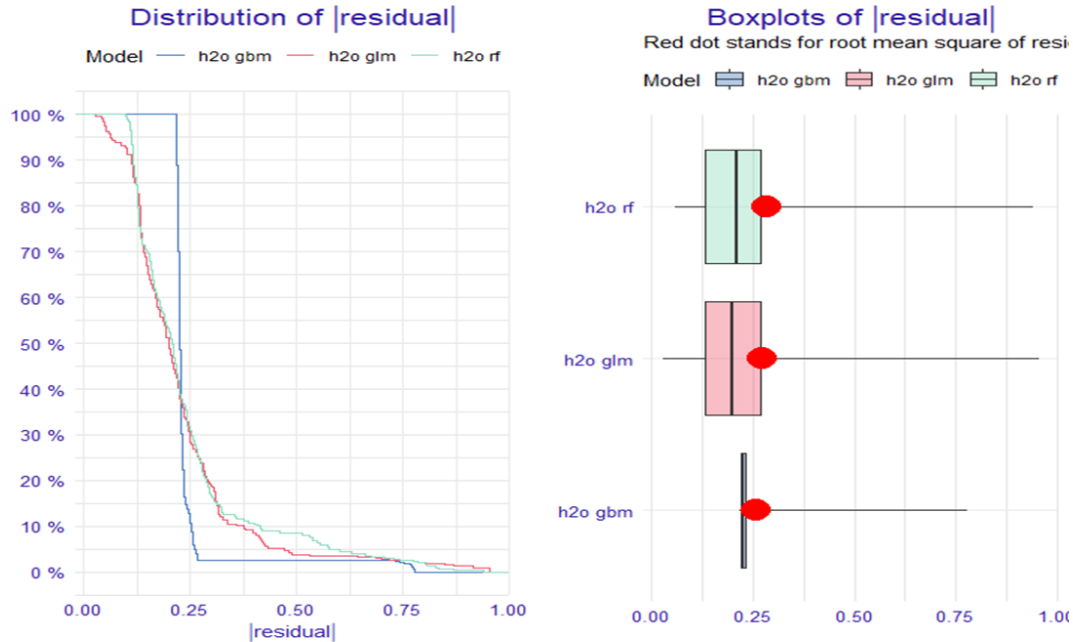
  

Distributed Random Forest (DRF)					
Predictores	Variable respuesta	Statistics	data_training	data_test	data_val
MONTO_CERTIFICADO	SINIESTRADO	MSE:	0.152	0.141	0.153
NUM_CUOTAS		RMSE:	0.390	0.375	0.391
diff_months		LogLoss:	0.476	0.445	0.475
N_Protestos		Mean Per-Class Error	0.339	0.287	0.340
TIPO_ESTRUCTURA		AUC:	0.714	0.768	0.693
DIAS_GRACIA		pr_auc:	0.486	0.557	0.379
Monto_Mora_Pesos		Gini:	0.427	0.536	0.386
		R^2:	0.136	0.206	0.079

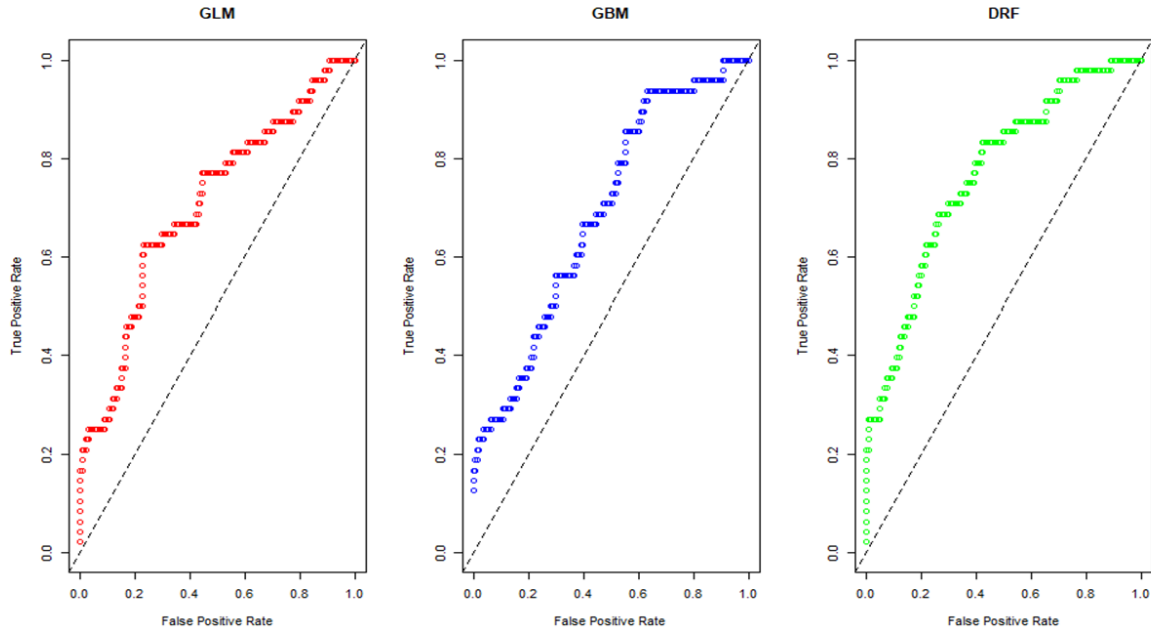
Como se ve en las tablas, los modelos presentan métricas similares en cuanto a la baja incertidumbre de las estimaciones en los tres conjuntos de datos. Si consideramos al AUC como una métrica de evaluación comparativa entre metodologías, en todos los modelos propuestos este valor no tiende a sobreajustar o producir un subajuste sobre los datos de test o validación, por lo que la generalización de los datos de entrenamiento no es traspasada a los otros conjuntos de datos. En relación al coeficiente de Gini, el modelo GLM es el que presenta un valor más cercano a 0 (0.443), lo que indica una mejor performance en comparación con los otros.



La siguiente figura presenta la distribución de los residuales en las predicciones de cada uno de los modelos. Si bien el modelo GBM no tiene una gran variabilidad en la distribución cuantil cuantil, en cada uno de los extremos la desviación con respecto al origen es mayor que la de los otros modelos, en cambio, el modelo DRF presenta una distribución más homogénea a través de los cuantiles. Si bien el modelo GBM no presenta colas pesadas en los residuales, la figura casi sin variación de la distribución a través de los cuantiles hace pensar que el modelo no es flexible para la predicción sobre otros conjuntos de datos. El gráfico boxplot de los mismos residuales en las metodologías propuestas muestra lo acotado de la distribución de los residuales en el modelo GBM, en cambio en el DRF y GLM, las medianas se encuentran casi en el mismo cuantil (0.25), pero con una mayor variación en el modelo DRF.



Las curvas ROC se presentan a continuación para cada uno de los modelos. Se aprecia que todos ellos sobrepasan el umbral del 0.5, lo que nos indica que la predicción no es en forma aleatoria, considerando a los predictores como explicativos en el comportamiento de SINIESTRADO2. El modelo que presenta la mejor predicción en terminos de la curva ROC (y AUC) es el DRF, con un área por sobre la recta que corta los ejes extremos mayor a los otros modelos evaluados en este problema.



Por último se presenta la matriz de confusión que evalúa la performance del modelo predictivo sobre un conjunto de datos ‘no visto’ (data\_test). De acuerdo a los verdaderos positivos y verdaderos negativos, el modelo que predice mejor estos valores es el GLM. Si nos fijamos en el Accuracy (Exactitud) para comparar los modelos, el GLM es el que presenta un valor mas cercano a 0 (0.737), lo que lo permite clasificar cerca del 73.4% a los clientes que efectivamente SINIESTRAN. El modelo que presenta una mejor predicción de los SINIESTRADOS es el DRF (recall = 0.688) y con un valor de especificidad de 0.739 (quienes NO SINIESTRAN). Además, considerando a la confusión (si el modelo predice clientes SINIESTRADOS, cuantos clasifica correctamente?), el modelo de mejor performance también es el GLM con un valor de 0.448.

GLM				GBM				DRF			
		Predicciones				Predicciones				Predicciones	
		No	Si			No	Si			No	Si
Observaciones	No	124	37	Observaciones	No	59	102	Observaciones	No	119	42
	Si	18	30		Si	3	45		Si	15	33

Modelos	Accuracy	Precision	Recall	Especificidad
GLM	0.737	0.448	0.625	0.770
GBM	0.498	0.306	0.938	0.366
DRF	0.727	0.440	0.688	0.739

## 5. CONCLUSIONES FINALES

- La unificación de la información disponible en las diferentes fuentes de información permitió la creación de dos modelos predictivos para predecir el SINIESTRO en clientes.
- Se crearon dos marcas para modelar y predecir a los clientes SINIESTRADOS. La primera marca, SINIESTRADO, esta en función de aquellos clientes que aparecen en la columna 'ESTADO\_CERTIFICADO' como 'Pagado Ach' o 'Siniestro parcial'. La segunda marca, SINIESTRADO2, esta en función de las mismas 'Pagado Ach' y 'Siniestro parcial' pero también considerando si alguna vez el cliente estuvo como Normalizado.
- Se identificaron las principales variables que permiten explicar el comportamiento de los clientes SINIESTRADOS mediante la 'selección de variables más importantes'. Lo anterior se hizo a través de diferentes metodologías de selección y que pueden encontrarse a disposición en los códigos adjuntos.
- Mediante 3 modelos de clasificación se obtuvo la probabilidad de que un cliente efectivamente pueda SINIESTRAR en función de los patrones encontrados en los modelos predictivos.
- Para la primera marca, SINIESTRADO, el modelo que presenta mejor performance es el GLM. El modelo GLM presenta mejores métricas de predicción basadas en la matriz de confusión, sobre todo en el valor de accuracy y en del recall, en donde sus valores son más altos que los dos modelos restantes. Si bien el DRF presenta valores de AUC estables para cada conjunto de datos utilizados, en términos predictivos el GLM obtiene una puntuación de 0.81 en el AUC mientras el DRF presenta un valor de 0.768 y el GBM de 0.713, lo que también puede apreciarse en el gráfico de residuales de las predicciones de cada uno de los modelos. Las medianas de los residuales en los 3 modelos son similares pero las colas pesadas en GBM y en el DRF son más pronunciadas que en el GLM.
- Para la segunda marca, SINIESTRADO2, el DRF es el que presenta una mejor performance predictiva. El DRF no presenta una generalización del conjunto de datos de entrenamiento (data\_training) con un valor alto del AUC, como sí en cambio lo hace el modelo GMB, el cual puede presentar esta generalización ya que va iterando en función del último paso del algoritmo. En relación a las predicciones y la matriz de confusión, el DRF presenta mejores predicciones en los clientes que efectivamente SINIESTRAN (recall = 0.688) y con una distribución de los residuales homogénea en las predicciones.
- Los resultados presentados anteriormente presentan una buena predicción sobre el conjunto de datos que se predicen ('data\_test'), ya que existe una alta incertidumbre propia del modelado estadístico, y además de la forma en que se abordó el problema. En futuros trabajos para modelar esta misma problemática, quizás sería interesante agregar un efecto aleatorio (variable latente) el cual incorporé en alguna medida el comportamiento dinámico del cliente, ya que se cuentan con observaciones en intervalos de tiempo específico pero no lo suficiente para incorporar una variabilidad temporal para cada uno de ellos. De esta forma se podría reducir la incertidumbre asociada al comportamiento de un mismo cliente durante el transcurso de un tiempo determinado.




## Anexo I

### Data Frame Summary

full\_join

Dimensions: 1361 x 3






Duplicates: 5

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	MONTO_CERTIFICADO [numeric]	Mean (sd) : 67900882 (125915851.9) min < med < max: 78 < 30431106 < 1668154163 IQR (CV) : 65104943 (1.9)	1221 distinct values		1361 (100%)	0 (0%)
2	NUM_CUOTAS [integer]	Mean (sd) : 23.8 (29.5) min < med < max: 1 < 9 < 144 IQR (CV) : 35 (1.2)	68 distinct values		1361 (100%)	0 (0%)
3	DIAS_GRACIA [integer]	Mean (sd) : 124.8 (155) min < med < max: 0 < 75.5 < 1634 IQR (CV) : 134 (1.2)	317 distinct values		1312 (96.4%)	49 (3.6%)

## Data Frame Summary

**Dimensions:** 1361 x 5

**Duplicates:** 971

Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
N_Moras [numeric]	Mean (sd) : 2.2 (11.7) min < med < max: 0 < 0 < 249 IQR (CV) : 1 (5.4)	40 distinct values		0 (0%)
Monto_Mora_Pesos [integer]	Mean (sd) : 6771010.4 (48213118.3) min < med < max: 0 < 0 < 850053262 IQR (CV) : 83600 (7.1)	365 distinct values		0 (0%)
N_Multas [numeric]	Mean (sd) : 0.2 (2) min < med < max: 0 < 0 < 69 IQR (CV) : 0 (13.5)	11 distinct values		0 (0%)
N_Protestos [numeric]	Mean (sd) : 1.5 (11) min < med < max: 0 < 0 < 298 IQR (CV) : 0 (7.3)	35 distinct values		0 (0%)
Monto_Protestos_Pesos [numeric]	Mean (sd) : 46894.4 (865864.4) min < med < max: 0 < 0 < 27204802 IQR (CV) : 0 (18.5)	13 distinct values		0 (0%)

## Anexo II

```
#=====
#                               Compara distribución
#=====
#                               Monto certificado
#=====
wilcox.test(full_join$MONTO_CERTIFICADO ~ SINIESTRADO, data=full_join)
wilcox.test(log(full_join$MONTO_CERTIFICADO) ~ SINIESTRADO, data=full_join)

# bin/discretize (cut continuous variable into equally sized groups)
(q<-quantile(full_join$MONTO_CERTIFICADO , seq(0, 1, .2)))

# MONTO CERTIFICADO
full_join$monto_bin <- cut(full_join$MONTO_CERTIFICADO, breaks=q, include.lowest=F)
summary(full_join$monto_bin)

(tab<-with(full_join, table(monto_bin, SINIESTRADO))) # counts

print(prop.table(tab,2)) # proportions

# tabulate the binned data
(q2<-quantile(log(full_join$MONTO_CERTIFICADO) , seq(0, 1, .2)))
full_join$monto_bin_log <- cut(log(full_join$MONTO_CERTIFICADO), breaks=q2, include.lowest=F)
summary(full_join$monto_bin_log)

(tab2<-with(full_join, table(monto_bin_log, SINIESTRADO))) # counts

print(prop.table(tab,2)) # proportions
print(prop.table(tab2,2)) # proportions

#=====
#                               NUM_CUOTAS
#=====
wilcox.test(full_join$NUM_CUOTAS ~ SINIESTRADO, data=full_join)
wilcox.test(log(full_join$NUM_CUOTAS) ~ SINIESTRADO, data=full_join)

# bin/discretize (cut continuous variable into equally sized groups)
breaks =unique(quantile(full_join$NUM_CUOTAS, probs = seq(0, 1, 0.2)))

# MONTO CERTIFICADO
full_join$cuotas_bin <- cut(full_join$NUM_CUOTAS, breaks= breaks, include.lowest=F)
summary(full_join$cuotas_bin)

(tab<-with(full_join, table(cuotas_bin, SINIESTRADO))) # counts
print(prop.table(tab,2)) # proportions

# tabulate the binned data
breaks2 =unique(quantile(log(full_join$NUM_CUOTAS), probs = seq(0, 1, 0.2)))
full_join$cuotas_bin_log <- cut(log(full_join$NUM_CUOTAS), breaks= breaks2, include.lowest=F)
summary(full_join$cuotas_bin_log)

(tab2<-with(full_join, table(cuotas_bin_log, SINIESTRADO))) # counts
```

```

print(prop.table(tab,2)) # proportions
print(prop.table(tab2,2)) # proportions

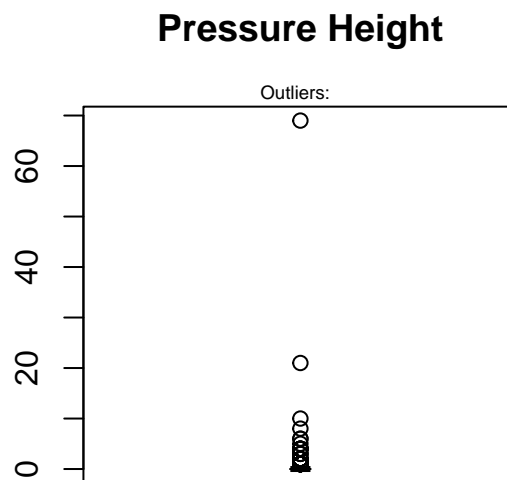
#####
#                               Detección de outliers
#####

# Outliers (Univariado)
N_multas_outliers <- boxplot.stats(full_join$N_multas)$out # outlier values.

Warning: Unknown or uninitialised column: 'N_multas'.

boxplot(full_join$N_Multas, main="Pressure Height", boxwex=0.1)
mtext(paste("Outliers: ", paste(N_multas_outliers, collapse=" ")), cex=0.6)

```



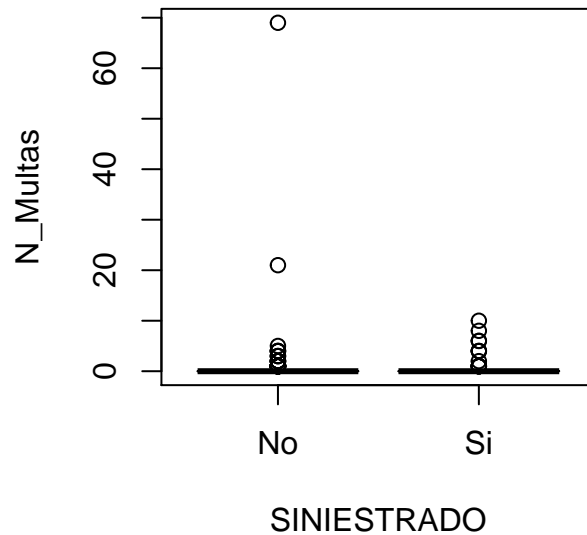
```

# Bivariado
boxplot(N_Multas ~ SINIESTRADO, data= full_join, main="N_multas vs SINIESTRADO") #

```

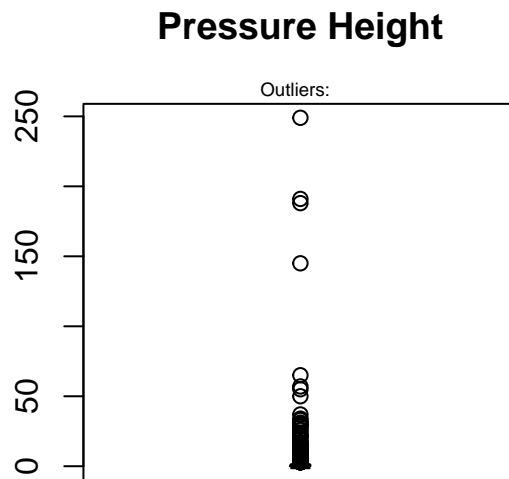


## N\_multas vs SINIESTRADO

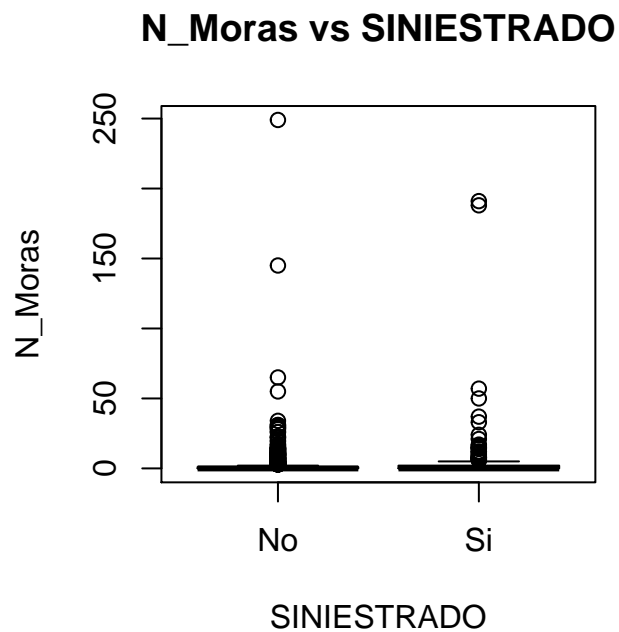


En N\_Multas quizás llamar ‘outliers’ a muchos de estos valores es inadecuado ya que casi todos los valores están concentrados en el 0, por lo que si algún cliente presenta 2 números de multas, ya es considerado un ‘outlier’. Los valores cercanos a 20 y 70 son lo que si pueden considerarse como ‘outliers’.

```
# Outliers (Univariado) ----> N_Moras
N_moras_outliers <- boxplot.stats(full_join$N_Moras)$out # outlier values.
boxplot(full_join$N_Moras, main="Pressure Height", boxwex=0.1)
mtext(paste("Outliers: ", paste(N_moras_outliers, collapse=", ")), cex=0.6)
```



```
# Bivariado
boxplot(N_Moras ~ SINIESTRADO, data= full_join, main="N_Moras vs SINIESTRADO") #
```



El número de moras también presenta valores ‘outliers’ si observamos un gráfico simple univariado de la distribución de sus observaciones. Estos valores deberían tener un tratamiento especial y ver si tienen una potencial implicancia en los resultados finales.