ESCUELA POLITÉCNICA NACIONAL

Facultad de Ciencias

Modelos de Riesgo

Tema: Modelo Logístico

Integrantes:
Luis Amagua
Nohely Córdova
Erika Ramírez

Quito - Ecuador
8 de enero de 2022

Índice

1. Introducción

El sistema bancario engloba a las instituciones que participan en una economía como punto de encuentro entre el ahorro y la inversión. Dicho de otra forma, los bancos se dedican a captar recursos a través de créditos y mediante la realización de inversiones (BBVA, 2021)[2]. Los créditos bancarios son montos financieros que los bancos ponen a disposición de ciertos clientes, mediante un acuerdo con determinadas condiciones de devolución. La posibilidad de que estos clientes incumplan el acuerdo es un riesgo permanente que los bancos asumen. Ahora bien, una gestión eficaz del conjunto de riesgos asumidos, es parte fundamental de la toma de decisiones y contribuye a la creación de valor (Soler et al., 1999, p. 1)[17].

El riesgo de crédito es la probabilidad de que, a su vencimiento, una entidad no haga frente, en parte o en su totalidad, a su obligación de devolver una deuda o rendimiento, acordado sobre el instrumento financiero (Chorafas, 2000, como se citó en Saavedra García Saavedra García, 2010)[6]. Una adecuada gestión del riesgo de crédito permite disminuir las posibilidades de sufrir pérdidas debido a una concesión de crédito no apropiada. En este sentido las instituciones financieras requieren anticiparse al comportamiento futuro de sus clientes con el fin de decidir si ofrecer sus productos y bajo qué condiciones. Es en torno a esta necesidad que surgen como herramientas los modelos predictivos para la construcción de scoring de créditos. "Los modelos tipo scoring son instrumentos de clasificación o puntuación utilizados por las entidades financieras en la decisión de otorgar un crédito. Esta metodología de medición de incumplimiento crediticio, tomó gran importancia desde el acuerdo sobre legislación y regulación bancaria emitido por el Comité de Supervisión Bancaria de Basilea en el año 2004 ... Esta probabilidad se puede estimar considerando las características del individuo y del crédito que éste solicita" (Moreno Valencia, 2013, p. 5)[14]

Para el presente trabajo, se propone crear un modelo credit scoring, que permite valorar de forma automática el riesgo asociado a cada solicitud de crédito. Riesgo que estará en función de características propias del cliente, además de variables macroeconómicas, que van a definir cada observación, es decir, cada solicitud de crédito.

2. Objetivos

- Identificar las variables que tienen influencia al decidir el otorgamiento de un crédito.
- Pronosticar el comportamiento de los individuos en torno al acuerdo del crédito.
- Construir un credit score que represente la probabilidad de que un cliente cumpla con el acuerdo del crédito.
- Determinar que variables ingresan en el modelo, seleccionando aquellas características con mayor poder predictivo aplicando métodos paramétricos o no paramétricos).
- Establecer grupos de riesgo homogéneos en las dos filosofías(PIT Y TTC), para analizar sus ciclos económicos y atribuir las probabilidades de incumplimiento que nos permitan dar una respuesta adecuada sobre si se otorga o no un crédito a una persona deacuerdo a las clasificaciones mencionadas.
- Validar los modelos, a través de técnicas como las curvas ROC.
- Calcular la Perdida Esperada (PE) en base de los datos obtenidos en los procesos previos para cada una de las filosofías (PIT y TTC)

3. Análisis de los datos

3.1. Limpieza de la base de datos

Previo a cualquier análisis, es necesario trabajar o depurar la base de datos con el fin de identificar errores en el registro de los atributos, porcentaje de datos faltantes, correcta asignación de clase, nomenclatura, unidad de medida y hasta errores tipográficos. Los datos faltantes, a diferencia de las demás consideraciones, necesitan un trato especial, ya que, dependiendo de su peso en la base, se tomarían distintas acciones (eliminarlos o imputarlos) con el fin de no perder información valiosa que pueda aportar al modelo. Una referencia para poder

eliminar una variable, es que los datos faltantes superen el 30 % de representación. Caso contrario, dependiendo de la distribución de los datos, se aplicaría un método básico de imputación con medidas de tendencia central, como la media (distribución simétrica) o la mediana (distribución asimétrica). Ventajosamente esta base de datos no tiene datos faltantes ni errores en el registro de la información.

3.2. Descripción de los datos

Se utilizará una base de datos ficticia de una institución financiera. Dicha base se compone de 24 variables y 22.965 observaciones[1], las cuales se registran desde el 29 de enero de 2018 hasta el 22 de marzo de 2019. Se dispone de información referente a si los clientes de esta institución financiera obtuvieron una marca de mora en sus créditos durante aproximadamente el periodo de un año y 2 meses calendario. Conjuntamente, la base de datos proporciona el historial de los clientes frente a otros productos financieros y variables sociales como edad, género e instrucción.

3.3. Descripción y clasificación de las variables

Variables categóricas

1. **GENERO:** Es el género de cada cliente al origen ("FEM","MAS")
2. **SUCURSAL:** Muestra la ciudad de origen del cliente (AMBATO, CUENCA, ESMERALDAS, GUAYAQUIL, IBARRA, LATACUNGA, LOJA, MACHALA, MANTA, QUEVEDO, QUITO, RIOBAMBA, SANTO DOMINGO)
3. **FORMA PAGO:** Indica si la tarjeta tiene débito automático para el pago (1= SI, 0 = NO)
4. **ORIGEN APROBACIÓN:** Indica el canal de origen de la tarjeta (0= Proactivo, 1= Demanda)
5. **MARCA CUENTA CORRIENTE:** Indica si el cliente tiene cuenta corriente activa (1 = SI, 0 = NO).
6. **MARCA CUENTA AHORROS:** Indica si el cliente tiene cuenta de ahorros activa (1 = SI, 0 = NO).
7. **INSTRUCCIÓN:** Es la instrucción educativa del cliente al origen (PRI, SEC, TEC, UNI)
8. **SEGMENTO RIESGO:** Es una calificación interna que tiene el cliente según el uso de los productos (A, B, C, D, E).

Variables numéricas

1. **CÓDIGO ID:** Es el código que identifica a la tarjeta.
2. **SALDO TOTAL TARJETA:** Es el saldo usado de la tarjeta al último corte.
3. **CUPO PROMEDIO TARJETA:** Es la línea de crédito promedio que tiene la tarjeta en los últimos 2 años.
4. **SALDO UTILIZ PROM CLIENTE:** Es el promedio del saldo usado en los últimos 6 meses.
5. **CANTIDAD TOTAL AVANCES:** Es el número de avances aprobados por ventanilla de la tarjeta.
6. **ANTIGUEDAD TARJETA ANIOS:** Antigüedad de la tarjeta en años.
7. **PROMEDIO MENSUAL CONSUMOS LOCALES:** Es la facturación promedio en establecimientos.
8. **MAXIMO NUM DIAS VENCIDO:** Es el máximo número de días vencidos que tuvo la tarjeta últimos 6 meses.

[1]La base proporciona también un apartado de 4 variables sistémicas con 60 observaciones.

9. **NUMERO OPERACIONES TITULAR:** Es el número de operaciones de crédito (excepto TC) que tiene el cliente vigente y cancelado en los últimos 2 años en el banco.
10. **PROMEDIO DIAS SOBREGIRO CC:** Indica el número máximo de días de sobregiro que tuvo el cliente en su cuenta corriente en los últimos 6 meses.
11. **PROMEDIO MENSUAL SALDO CUENTA PASIVO:** Muestra los saldos promedios en productos el pasivo en los últimos 6 meses.
12. **RIESGO CLIENTE TOTAL GFP:** Muestra la deuda total del cliente incluyendo otros créditos o productos.
13. **VALOR DEPOSITO A PLAZO:** Muestra el valor en pólizas de acumulación al corte.
14. **EDAD:** Es la edad en años del cliente al corte de datos de la información.
15. **NUM TC SIST FIN:** Número de TC que el cliente tiene en el sistema financiero (sin incluir Produbanco).

Variables Macroeconómicas

1. **ICC_Indice confianza consumidor:** Indicador económico que mide el grado de optimismo que los consumidore tienen sobre el estado general de la economía y sobre su situación financiera personal. Además permite saber la disposición que muestra la ciudadanía de consumir en la economía.
2. **IDEAC:** Indicador de la tendencia que seguiría la producción nacional. Permite describir los cambios en el volumen de la actividad económica del país, con periodicidad mensual.
3. **CRUDO ORIENTE:** Precio del crudo oriente en Ecuador.
4. **PETRÓLEO WTI:** Precio del petróleo crudo WTI.

3.4. Variable dependiente

La variable dependiente Y($MARCAMORA\_TARJETA$) es una variable dicotómica que muestra si una tarjeta llegó a tener la marca de mora durante un periodo de tiempo t, este tiempo t esta considerado desde el 29 de enero de 2018 hasta el 22 de marzo de 2019. Las categorías son:

$$Y = \begin{cases} 1 & \text{si el cliente tiene un crédito en mora (mal pagador)} \\ 0 & \text{No tiene Mora (buen pagador)} \end{cases}$$

A continuación se muestra en una tabla, la cantidad de observaciones de 0 y 1, además del porcentaje correspondiente en la variable dependiente Y.

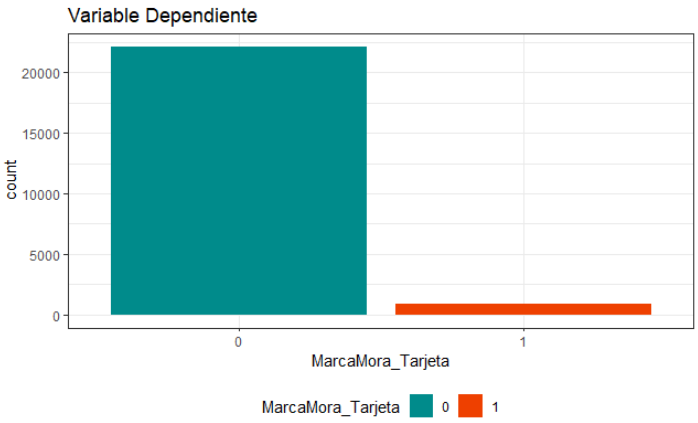


Figura 1: Cantidad de 0 y 1 en la variable Y

Y	Cantidad	Porcentaje
0	22.100	0,96
1	865	0,04

Tabla 1: Variable dependiente

En la Tabla 1 podemos observar que el 96 % de personas no tienen mora; es decir, son buenos pagadores y el 4 % tienen mora, lo que significa que son malos pagadores.

4. Balanceo de la data

Como se puede observar en el previo análisis se tiene un caso de datos de eventos raros, ya que existen muchos "no eventos"(0 = 96 %) frente a los "si eventos"(1 = 4 %) de la variable dependiente Marca_mora_tarjeta. Dicho de otra forma, los datos no están balanceados porque la probabilidad de ocurrencia de un suceso es muy baja. Este problema, complica la descripción y predicción que podría ser obtenida de esta base de datos. Específicamente, una regresión logística puede subestimar drásticamente la probabilidad de eventos raros. Un procedimiento útil para tratar este tipo de información es el balanceo de los datos. Este método sirve para compensar las diferencias en la muestra y en las fracciones de población de las inducidas por muestreo basado en elecciones (King et al, 2001)[13].

Para balancear se puede aplicar el método de sobremuestreo (oversampling), submuestreo (undersampling) o ponderación (weighting). En este caso de un modelo logístico para evaluar el riesgo crediticio, el interés recae en predecir si un cliente es buen o mal pagador. La variable dependiente Marca_mora_tarjeta registra apenas un 4 % de malos pagadores, o de personas que registran una marca de mora en su tarjeta. En caso de aplicar un sobremuestreo se podría obtener una muestra mayor, manteniendo los buenos pagadores como 80 % de la base de datos y replicando el 4 % original de malos pagadores, hasta llegar a un 20 %. Caso contrario, de aplicarse el submuestreo se obtendría una muestra menor, en la que el 4 % original de malos pagadores constituya el 20 % de la muestre, reduciento el número de buenos pagadores seleccionados. La tercera opción es la ponderación, bajo este método, los individuos no se desechan ni se replican. Se balancea el conjunto de datos (20 % - 80 %) asignando un peso en función de si es mal pagador (Marca_mora_tarjeta=1) o si es buen pagador (Marca_mora_tarjeta=0) y ponderando por esta variable al momento de estimar el modelo. En regresiones logísticas es común aplicar una proporción 20-80 (Hernández, 2019)[11].

Para hacer el balanceo de la data se va a tomar lo siguiente:

Y	Cantidad	Porcentaje
0	18.409	0,80
1	4.556	0,20

A continuación se muestra en una tabla, la cantidad de observaciones de 0 y 1, con la data balanceada

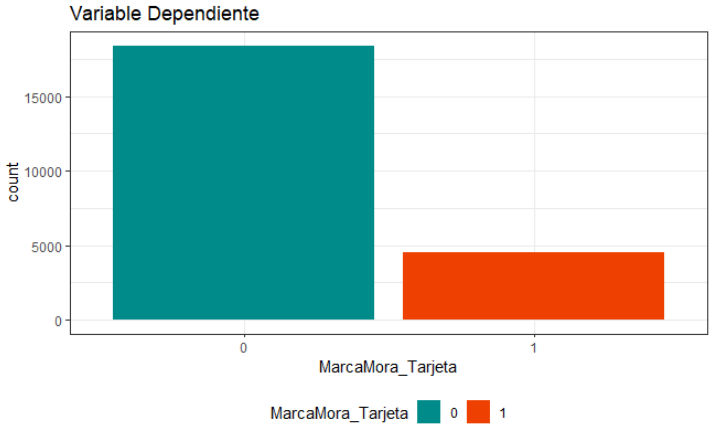


Figura 2: Cantidad de 0 y 1 en la variable Y

5. Muestra de modelamiento y validación

Para el trabajo se ha utilizado el software estadístico R. Primero se hará una lectura de los datos y se comprobará que el tipo de variables sea el correcto; es decir, si variables son numéricas o categóricas.

A continuación se divide a la data en dos submuestras aleatorias:

- Train: Para el desarollo del modelo (modelización)
- Test: Para la validación del modelo (validación)

Se han tomado las submuestras de tal forma que mantengan la misma distribución de la variable dependiente, en donde se ha tomado el 75 % de la data original para Train y el 25 % para Test.

6. Análisis descriptivo

Para el análisis descriptivo se toma la data Train, y se realiza la estadística descriptiva tanto para las variables numéricas como para las categóricas. Además, se identifica los valores atípicos, con la finalidad de que no ocasionen un efecto desproporcionado en los resultados estadísticos.

6.1. Variables Numéricas

A continuación se muestra un resumen de los datos:

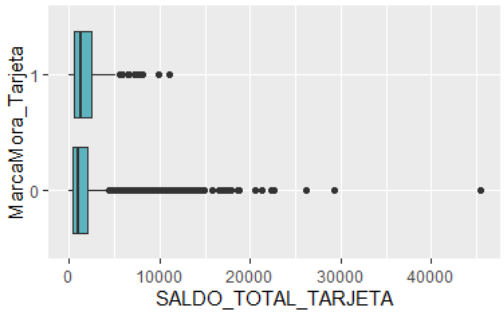
Variables Numéricas	Min.	1st Q.	Median	Mean	3rd Q.	Max.	Sd
SALDO_TOTAL_TARJETA	0.0	461.9	1014.7	1547.2	2065.9	45392.8	1778.06
CUPO_PROMEDIO_TARJETA	0	900	2200	2820	3000	50000	3096.98
SALDO_UTILIZ_PROM_CLIENTE	0.0	340.3	854.1	1318.2	1822.0	32850.1	1585.07
CANTIDAD_TOTAL_AVANCES	0.000	0.000	0.000	1.076	0.000	79.000	2.93
ANTIGUEDAD_TARJETA_ANIOS	0.000	2.000	3.000	4.272	7.000	12.000	3.39
PROMEDIO_MENSUAL_CONSUMOS_LOCALES	0.00	0.00	42.95	201.41	191.66	35956.00	657.13
MAXIMO_NUM_DIAS_VENCIDO	0.00	0.00	4.00	10.84	19.00	85.00	14.29
NUMERO_OPERACIONES_TITULAR	0.000	2.000	3.000	3.138	4.000	12.000	1.71
PROMEDIO_DIAS_SOBREGIRO_CC	0.0000	0.0000	0.0000	0.6017	0.0000	128.0000	3.08
PROMEDIO_MENSUAL_SALDO_CUENTA_PASIVO	-58423.4	0.0	80.4	2649.3	802.1	873976.0	17616.16
RIESGO_CLIENTE_TOTAL_GFP	0.0	523.4	1235.5	4292.8	2594.4	1715176.4	22832.29
VALOR_DEPOSITO_A_PLAZO	0	0	0	1820	0	2281686	31178.79
EDAD	23.00	29.00	36.00	36.49	44.00	50.00	8.08
NUM_TC_SIST_FIM	0.000	1.000	2.000	2.492	4.000	5.000	1.70

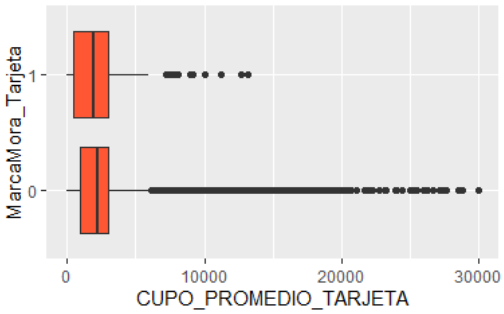
Tabla 2: Estadística Básica de variables cuantitativas

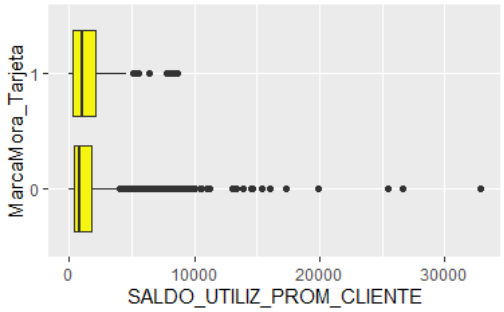
En la tabla 2 se puede apreciar que las variables saldo_total_tarjeta, cupo_promedio_tarjeta, saldo_utiliz_prom_cliente, promedio_mensual_saldo_cuenta_pasivo, riesgo_cliente_total_gfp y valor_deposito_a_plazo tienen una alta desviación estándar, lo que indica que la distribución de sus datos no se agrupa cerca de su media. Dada la naturaleza de estas variables, esto resulta coherente ya que responde a la diversidad de clientes y su diferente uso de los productos financieros.

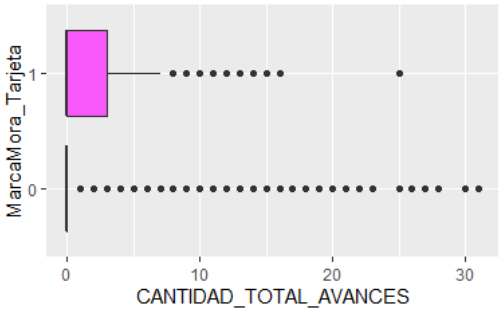
Para las variables numero_operaciones_titular, edad y num_tc_sist_fim, se presume que su distribución es simétrica, ya que la media y la mediana son muy cercanas. Otro tipo de información que se puede extraer de la tabla 2 es que los clientes de esta institución financiera oscilan entre los 23 y los 50 años y en su mayoría tienen alrededor de 36 años.

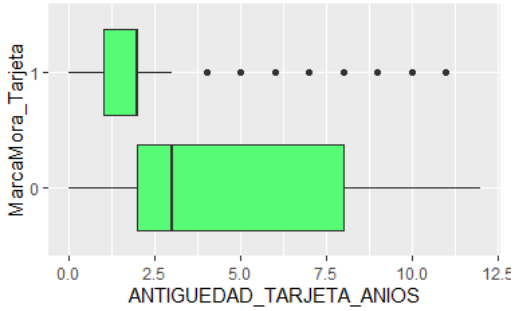
De la Tabla 2 se puede observar que probablemente existan valores atípicos, lo cual se confirmará mediante diagramas de caja, considerando los valores para 0 y 1 de variable dependiente $MarcaMora\_Targeta$

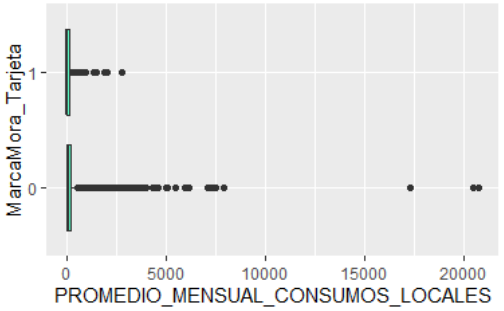


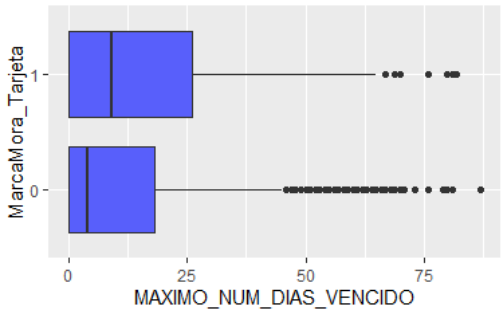


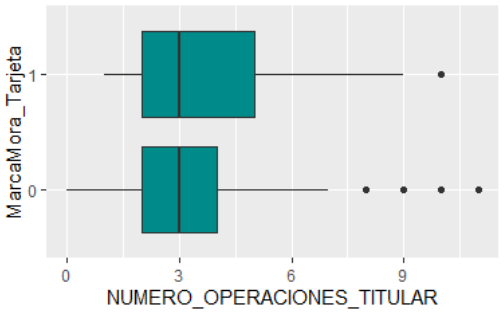


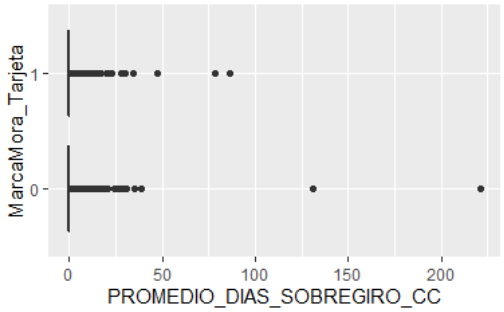


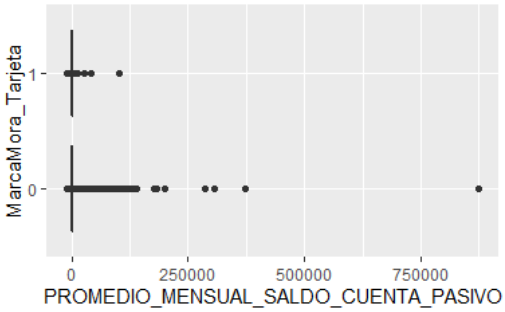


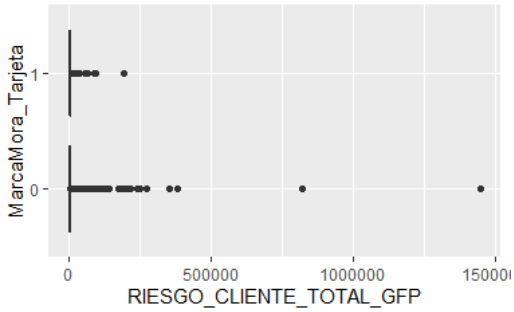


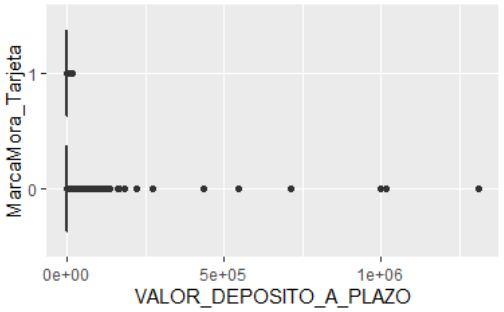

MarcaMora_Tarjeta
1
0
0
25
50
75
MAXIMO_NUM_DIAS_VENCIDO

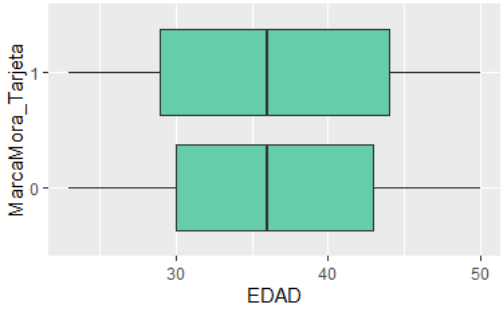

MarcaMora_Tarjeta
1
0
0
3
6
9
NUMERO_OPERACIONES_TITULAR

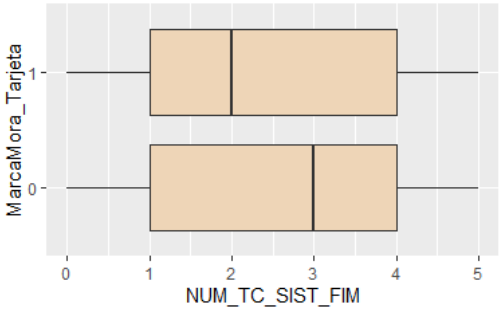

MarcaMora_Tarjeta
1
0
0
50
100
150
200
PROMEDIO_DIAS_SOBREGIRO_CC

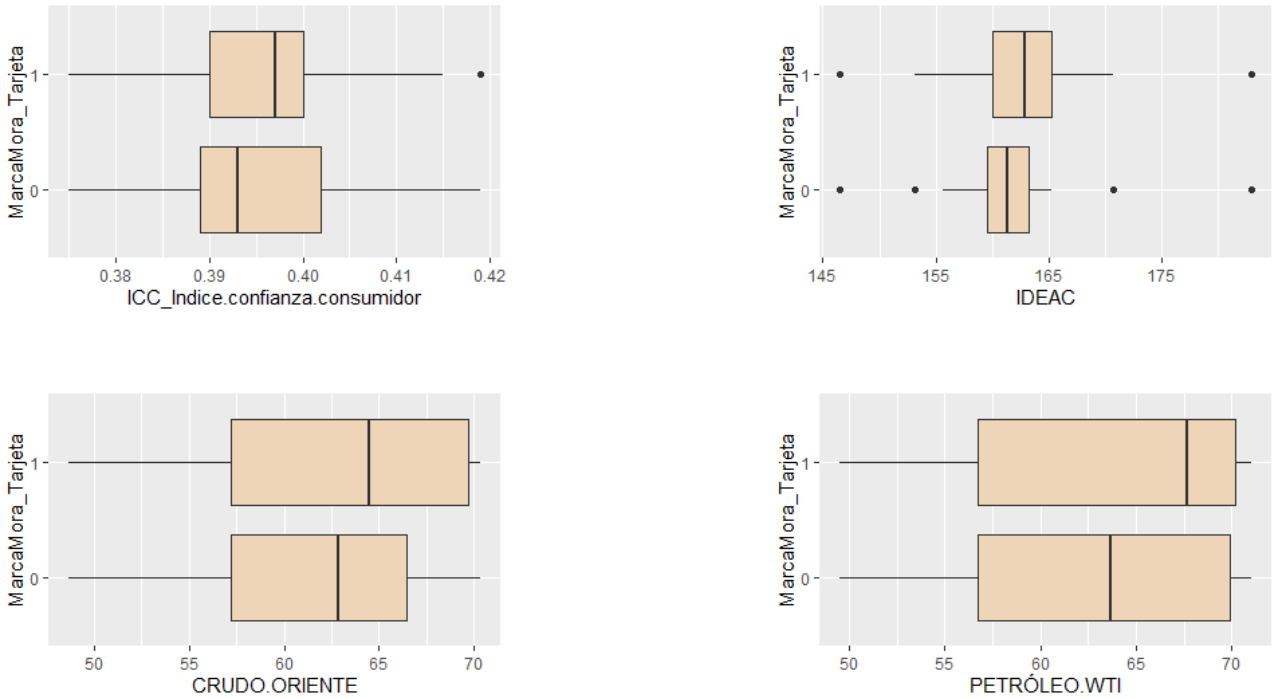

MarcaMora_Tarjeta
1
0
0
250000
500000
750000
PROMEDIO_MENSUAL_SALDO_CUENTA_PASIVO


MarcaMora_Tarjeta
1
0
0
500000
1000000
RIESGO_CLIENTE_TOTAL_GFP


MarcaMora_Tarjeta
1
0
0e+00
5e+05
1e+06
VALOR_DEPOSITO_A_PLAZO

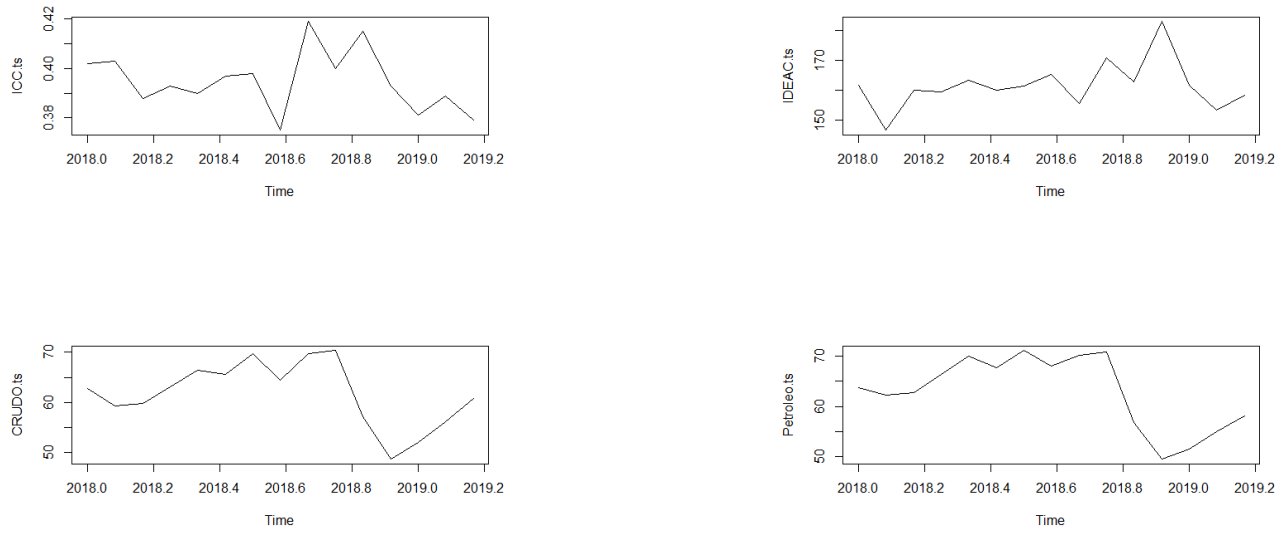

MarcaMora_Tarjeta
1
0
30
40
50
EDAD


MarcaMora_Tarjeta
1
0
0
1
2
3
4
5
NUM_TC_SIST_FIM



Se puede observar que existe una importante presencia de valores atípicos en gran parte de las variables. Por otro lado, variables como: "Antigüedad_tarjeta_años", "Número_operaciones_titular", "edad" y "Num_tc_sist_fim", no evidenciaron una importante cantidad de valores fuera del rango intercuartil.

En cuanto a las variables numéricas de carácter macroeconómico, a continuación se presenta el gráfico de su serie temporal:



A simple vista no se puede observar un patrón de comportamiento en las series de tiempo de las variables ICC, IDEAC, CRUDO y Petróleo WTI. Para comprobar la presencia de estacionariedad, se procede a aplicar la prueba de raíz unitaria de Dickey Fuller Aumentada. Este test tiene la hipótesis nula de existencia de raíz unitaria. Como se puede comprobar en los siguientes cuadros, todas las variables sistémicas tienen estacionariedad.

Augmented Dickey-Fuller Unit Root Test on SER01

Null Hypothesis: SER01 has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on AIC, maxlag=43)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-131.3834	0.0001
Test critical values:	1% level	-3.430558	
	5% level	-2.861516	
	10% level	-2.566798	

*MacKinnon (1996) one-sided p-values.

ICC_ts

Augmented Dickey-Fuller Unit Root Test on SER02

Null Hypothesis: SER02 has a unit root
Exogenous: Constant
Lag Length: 1 (Automatic - based on AIC, maxlag=43)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-90.68718	0.0001
Test critical values:	1% level	-3.430558	
	5% level	-2.861516	
	10% level	-2.566798	

*MacKinnon (1996) one-sided p-values.

IDEAC_ts

Augmented Dickey-Fuller Unit Root Test on SER03

Null Hypothesis: SER03 has a unit root
Exogenous: Constant
Lag Length: 5 (Automatic - based on AIC, maxlag=43)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-52.59036	0.0001
Test critical values:	1% level	-3.430558	
	5% level	-2.861516	
	10% level	-2.566798	

*MacKinnon (1996) one-sided p-values.

Crudo_ts

Augmented Dickey-Fuller Unit Root Test on SER04

Null Hypothesis: SER04 has a unit root
Exogenous: Constant
Lag Length: 5 (Automatic - based on AIC, maxlag=43)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-52.52206	0.0001
Test critical values:	1% level	-3.430558	
	5% level	-2.861516	
	10% level	-2.566798	

*MacKinnon (1996) one-sided p-values.

Petroleo_ts

En la figura 3 se verificará si existe correlación entre las variables numéricas, tanto idiosincráticas como sistémicas. Como se observa que las variables no están fuertemente correlacionadas, se puede inferir que no existe multicolinealidad.

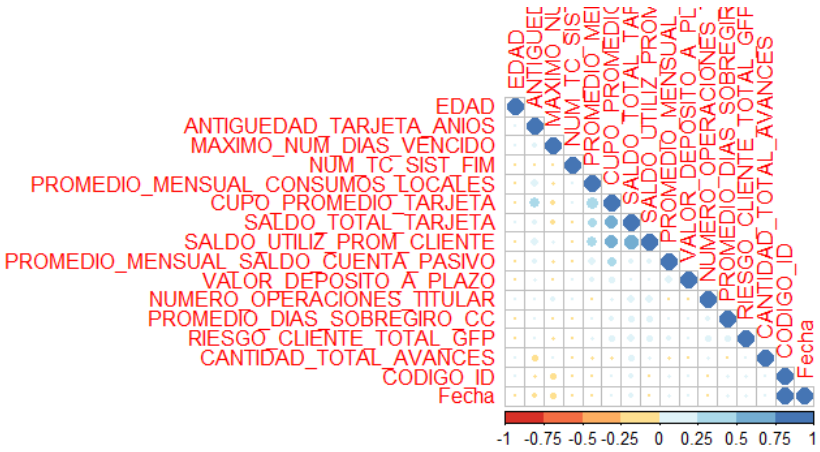
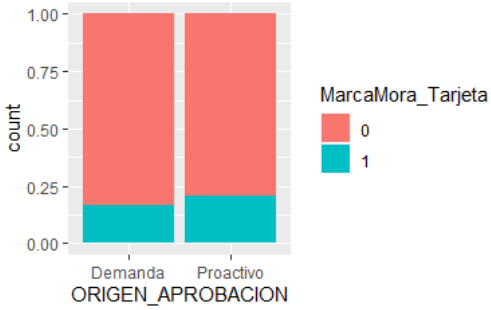


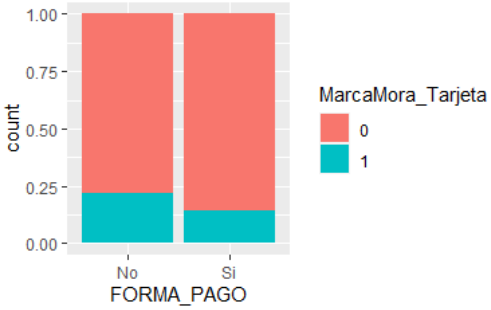
Figura 3: correlación entre variables

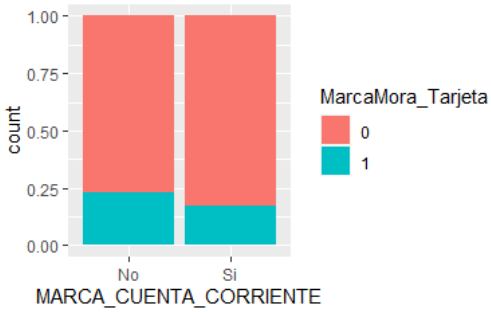
Para nuestro estudio, consideramos categorizar a la variable **Máximo_Número_días_vencido**, debido a que la mayor cantidad de datos se encuentra concentrada cerca del cero; por lo tanto, para un mejor análisis, se realizó la transformación a una variable dicotómica . Para visualizar las categorías se ha aplicado el método del codo, que lo explicaremos en la siguiente sección.

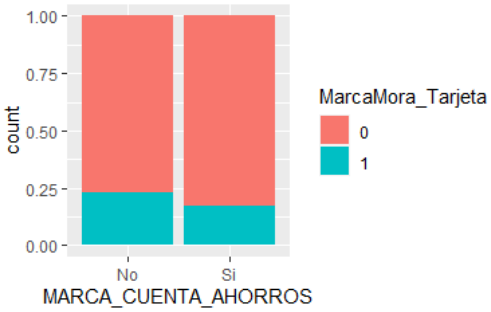
6.2. Variables Categóricas

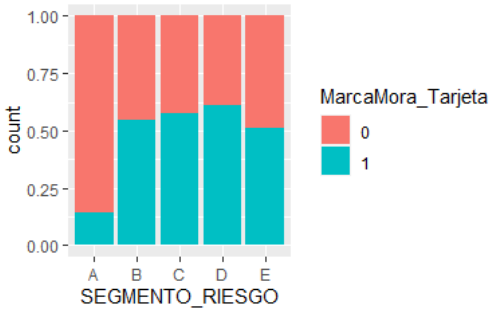
Para visualizar la distribución de los datos de las variables categóricas, se procedió a realizar gráficos de barras. Cabe destacar que como la base de datos fue balanceada, esta visualización de datos resulta interesante. Como por ejemplo en las variables segmento riesgo, donde se ve un claro crecimiento de la distribución de los malos pagadores conforme aumenta el segmento del riesgo. Así como también, de la variable sucursal, que presenta importantes variaciones de la distribución de la información de los malos pagadores por sucursal.

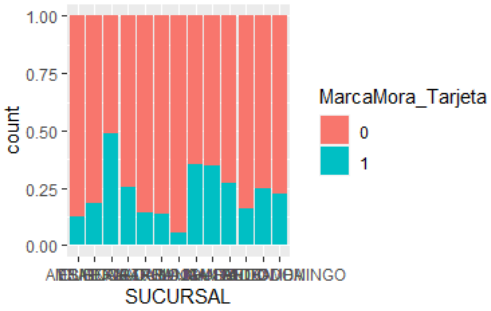

1.00
0.75
0.50
0.25
0.00
count
Demanda
Proactivo
ORIGEN_APROBACION
MarcaMora_Tarjeta
0
1

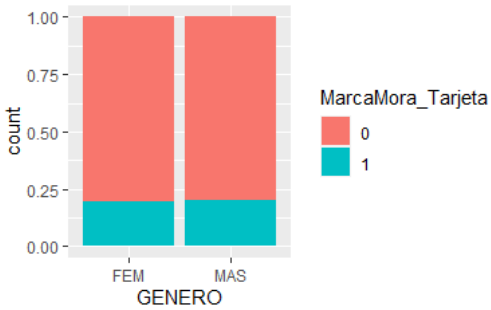

1.00
0.75
0.50
0.25
0.00
count
No
Si
FORMA_PAGO
MarcaMora_Tarjeta
0
1

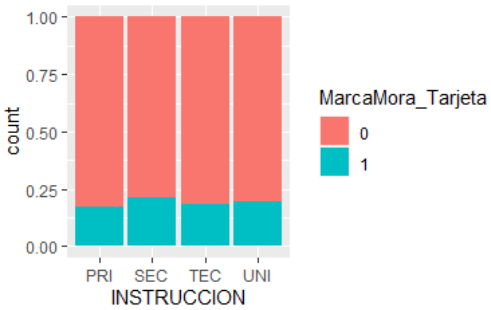

1.00
0.75
0.50
0.25
0.00
count
No
Si
MARCA_CUENTA_CORRIENTE
MarcaMora_Tarjeta
0
1

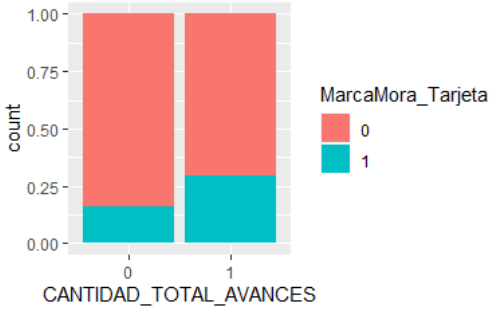

1.00
0.75
0.50
0.25
0.00
count
No
Si
MARCA_CUENTA_AHORROS
MarcaMora_Tarjeta
0
1

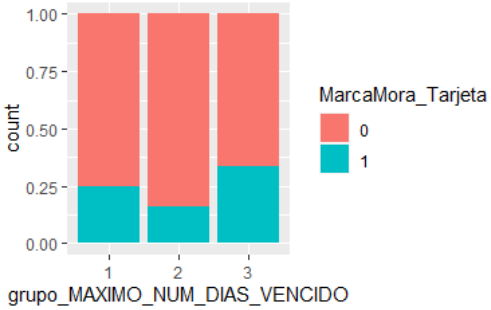

1.00
0.75
0.50
0.25
0.00
count
A
B
C
D
E
SEGMENTO_RIESGO
MarcaMora_Tarjeta
0
1


1.00
0.75
0.50
0.25
0.00
count
SUCURSAL
MarcaMora_Tarjeta
0
1


1.00
0.75
0.50
0.25
0.00
count
FEM
MAS
GENERO
MarcaMora_Tarjeta
0
1


1.00
0.75
0.50
0.25
0.00
count
PRI
SEC
TEC
UNI
INSTRUCCION
MarcaMora_Tarjeta
0
1





A continuación se presentan las tablas cruzadas de las variables categóricas y la variable dependiente $y = MarcaMora\_Tarjeta$, importante análisis para evaluar la representatividad de las categorías.

INSTRUCCION	0	Porcentaje	1	Porcentaje	Total	Porcentaje
PRI	3125	18.1 %	91	0.53 %	3216	18.7 %
SEC	9266	53.8 %	392	2.28 %	9658	56.1 %
TEC	332	1.9 %	11	0.06 %	343	2.0 %
UNI	3875	22.5 %	131	0.76 %	4006	23.3 %
Total	16598	96.4 %	625	3.63 %	17223	100.0 %

GENERO	0	Porcentaje	1	Porcentaje	Total	Porcentaje
FEM	6936	40.3 %	266	1.5 %	7202	41.8 %
MAS	9662	56.1 %	359	2.1 %	10021	58.2 %
Total	16598	96.4 %	625	3.6 %	17223	100.0 %

FORMA_PAGO	0	Porcentaje	1	Porcentaje	Total	Porcentaje
No	12350	71.7 %	517	3.0 %	12867	74.7 %
Si	4248	24.7 %	108	0.6 %	4356	25.3 %
Total	16598	96.4 %	625	3.6 %	17223	100.0 %

ORIGEN_APROBACION	0	Porcentaje	1	Porcentaje	Total	Porcentaje
Demanda	3659	21.2 %	119	0.7 %	3778	21.9 %
Proactivo	12939	75.1 %	506	2.9 %	13445	78.1 %
Total	16598	96.4 %	625	3.6 %	17223	100.0 %

MARCA_CUENTA_AHORROS	0	Porcentaje	1	Porcentaje	Total	Porcentaje
No	7885	45.8 %	353	2.0 %	8238	47.8 %
Si	8713	50.6 %	272	1.6 %	8985	52.2 %
Total	16598	96.4 %	625	3.6 %	17223	100.0 %

MARCA_CUENTA_CORRIENTE	0	Porcentaje	1	Porcentaje	Total	Porcentaje
No	7885	45.8 %	353	2.0 %	8238	47.8 %
Si	8713	50.6 %	272	1.6 %	8985	52.2 %
Total	16598	96.4 %	625	3.6 %	17223	100.0 %

SEGMENTO_RIESGO	0	Porcentaje	1	Porcentaje	Total	Porcentaje
A	15239	88.5 %	385	2.2 %	15624	90.7 %
B	817	4.7 %	139	0.8 %	956	5.6 %
C	148	0.9 %	31	0.2 %	179	1.0 %
D	128	0.7 %	32	0.2 %	160	0.9 %
E	266	1.5 %	38	0.2 %	304	1.8 %
Total	16598	96.4 %	625	3.6 %	17223	100.0 %

SUCURSAL	0	Porcentaje	1	Porcentaje	Total	Porcentaje
AMBATO	653	3.8 %	17	0.10 %	670	3.9 %
CUENCA	587	3.4 %	21	0.12 %	608	3.5 %
ESMERALDAS	115	0.7 %	13	0.08 %	128	0.7 %
GUAYAQUIL	4774	27.7 %	253	1.47 %	5027	29.2 %
IBARRA	363	2.1 %	11	0.06 %	374	2.2 %
LATACUNGA	64	0.4 %	3	0.02 %	67	0.4 %
LOJA	118	0.7 %	3	0.02 %	121	0.7 %
MACHALA	17	0.1 %	0	0.00 %	17	0.1 %
MANTA	336	2.0 %	20	0.12 %	356	2.1 %
QUEVEDO	115	0.7 %	6	0.03 %	121	0.7 %
QUITO	9152	53.1 %	265	1.54 %	9417	54.7 %
RIOBAMBA	102	0.6 %	4	0.02 %	106	0.6 %
SANTO DOMINGO	202	1.2 %	9	0.05 %	211	1.2 %
Total	16598	96.4 %	625	3.63 %	17223	100.0 %

Algunas variables (instrucción, segmento_riesgo y sucursal) poseen categorías poco representadas, por lo que éstas se agruparan y así se reducirá el número de categorías. Una representatividad pequeña tiende a tener una varianza muy grande por lo que la incertidumbre sería mayor. Dado que se busca obtener un modelo robusto, es importante asegurarse que cada categoría de las variables sea representativa[2].

7. Tratamiento de los datos

En los diagramas de caja presentados en la sección anterior, para las variables numéricas, pudimos observar que en su mayoría presentan datos atípicos. Es por esto, que es necesario tratar estos datos, con el fin de obtener el mejor modelo posible.

7.1. Recategorización de variables numéricas

Debido que tenemos variables que presentan datos atípicos, una forma de tratar estos datos es reemplazarlos por la media o una tranformacíon puesto que mantenerlos así, nos limitan para obtener un modelo adecuado. En primera instancia, trabajamos con el rango intercuartílico y transformación de datos por el método de Box Cox, pero al percatarnos que exiten variables que poseen pocos datos atípicos, la mejor decición fue recategorizarlos.

Como ejemplo podemos ver la variable: *Promedio_Mensual_Consumo_Locales*

[2]Es aconsejable que cada categoría tengo al menos un 5 % de representación.

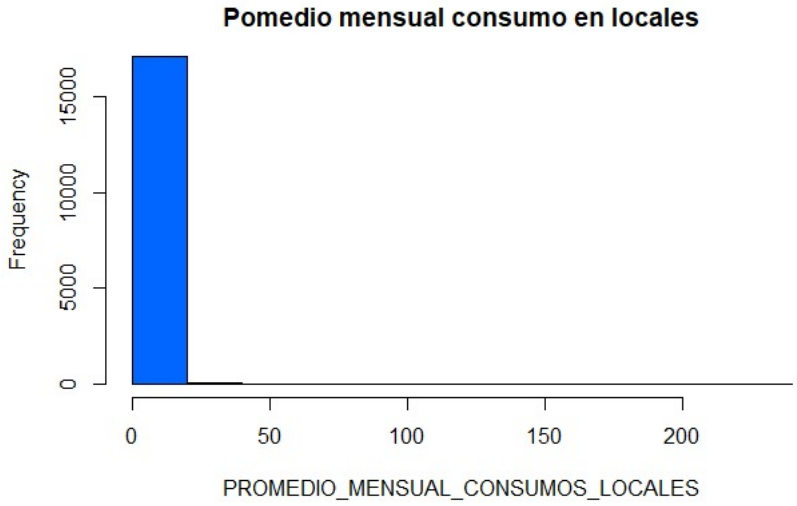


Figura 4: Promedio_Mensual_Consumo_Locales

Como se puede observar, en este caso no resulta conveniente categorizar; ya que la cantidad de datos atípicos es mínima.

Otro de los problemas que se presentó en la construcción del modelo es la aplicación de los métodos de Clustering (K-means o EM), específicamente la elección correcta del número de grupos. Existe una gran variedad de métodos que permiten elegir un número apropiado de Clusters para agrupar datos; como por ejemplo, el método del codo (elbow method), el criterio de Calinsky, el Affinity Propagation (AP), el Gap (también con su versión estadística), Dendrogramas, etc. Para este modelo se ha elegido trabajar con el método del codo, el mismo que servirá en distintas partes del trabajo para hallar el número óptimo de clusters, lo que se traduce en este caso de categorías para las variables numéricas.

7.1.1. Método del Codo

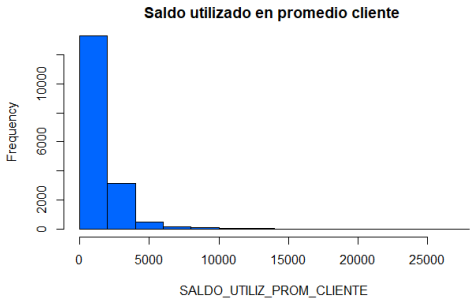
Un paso fundamental para cualquier algoritmo no supervisado es determinar el número óptimo de agrupaciones en las que se pueden agrupar los datos. El método del codo es uno de los métodos más populares para determinar este valor óptimo de k. El método consiste en graficar la variación explicada en función del número de conglomerados y elegir el codo de la curva como el número de clusters a utilizar. Este método utiliza los valores de la inercia obtenidos tras aplicar el K-means a diferente número de Clusters (desde 1 a n Clusters). Posteriormente se representa en una gráfica lineal la inercia respecto del número de Clusters. La inercia se define de la siguiente manera (Suma de las distancias al cuadrado de cada objeto del clúster a su centroide):

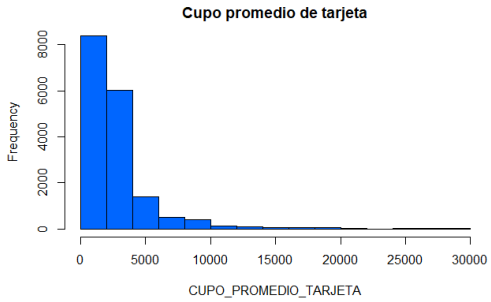
$$Inercia = \sum_{i=0}^{N} \|x_i - \mu\|^2$$

En teoría, en la gráfica lineal de la inercia se debería de apreciar un cambio brusco en su evolución, obteniendo en la línea representada una forma similar a la de un brazo y su codo.

1. **Creación variable:** *relacion_saldo_cupo*

   En el siguiente gráfico de frecuencias se puede observar a la distribución de las variables saldo utilizado en promedio clientes y cupo promedio de tarjeta:





En vista de la relación que tienen estas dos variables y del aporte de información que juntas podrían proporcionar, se procede a crear una nueva variable llamada *relacion_saldo_cupo*. En el siguiente diagrama de cajas se puede verificar que esta nueva variable presenta menos datos atípicos que las variables de forma individual; las cuales pasarán a ser descartadas de este modelo.

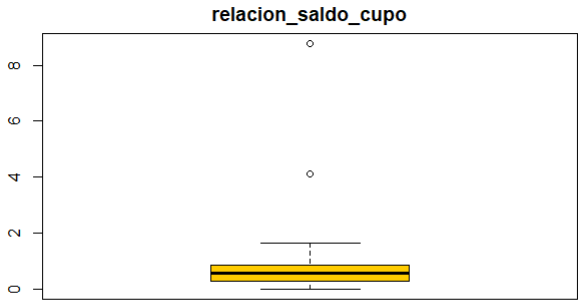


Figura 5: relación variables saldo/cupo

2. **Categorización variable:** Cantidad total de avances

Al observar las frecuencias de la variable cantidad total de avances, se decide categorizar esta variable, esperando que consiga mayor significancia en el modelo.

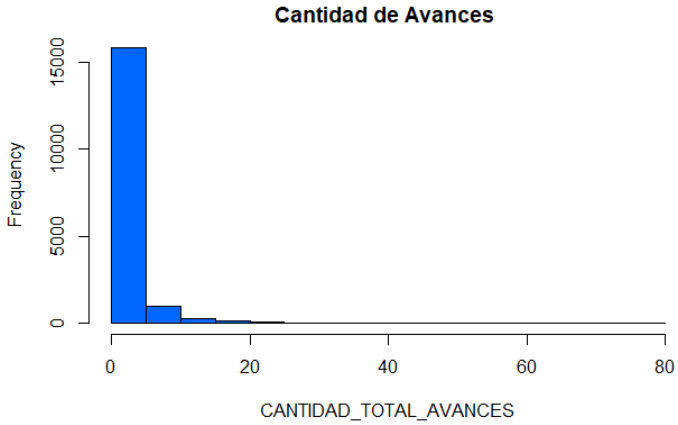


Figura 6: histograma: $CANTIDAD\_TOTAL\_AVANCES$

En la categorización se conservaron los valores de 0 y se les asignó esta misma etiqueta, lo cual representa

el número de avances aprobados por ventanilla. Los valores restantes, mayores a cero, fueron en cambio etiquetados bajo la categoría de 1.

3. **Categorización variable:** Máximo número de días vencidos

   Se tiene el histograma de la variable numérica $MAXIMO\_NUM\_DIAS\_VENCIDO$:

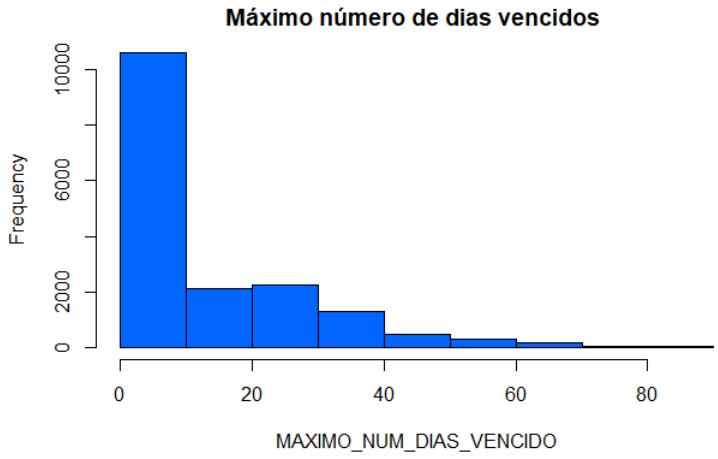


Figura 7: histograma:$MAXIMO\_NUM\_DIAS\_VENCIDO$

Se puede observar que en la variable **Máximo_Número_días_vencido**, la mayor cantidad de datos se concentran cerca del cero; por lo tanto, para un mejor análisis se procede a categorizar la variable. Para visualizar el número de categorías que se va a utilizar se ha aplicado el **Método del Codo**.

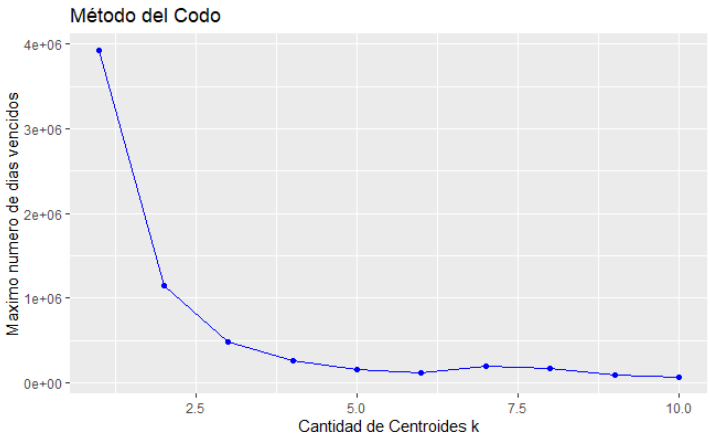


Figura 8: Categorización de Variable

Se puede observar que este método sugiere tomar entre 3 y 4 grupos. En este sentido se ha decidido tomar 4 grupos para una mejor representación. A continuación se muestra la distribución de frecuencias de los 4 grupos:

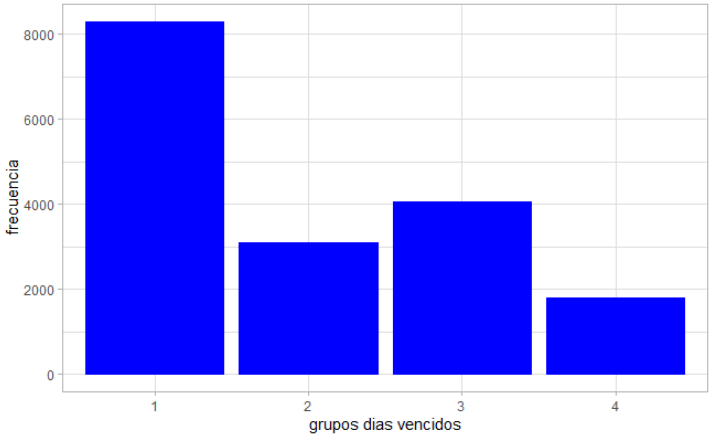


Figura 9: Frecuencia de los grupos

El intervalo del número de días vencidos que considerará cada grupo se clasifica de la siguiente forma:

- Grupo 1: [0,4]
- Grupo 2: [5,15]
- Grupo 3: [16, 33]
- Grupo 4: [34, 87]

4. Categorización variable: Edad

   Al analizar la variable edad, respecto al número total de clientes se tiene lo siguiente:

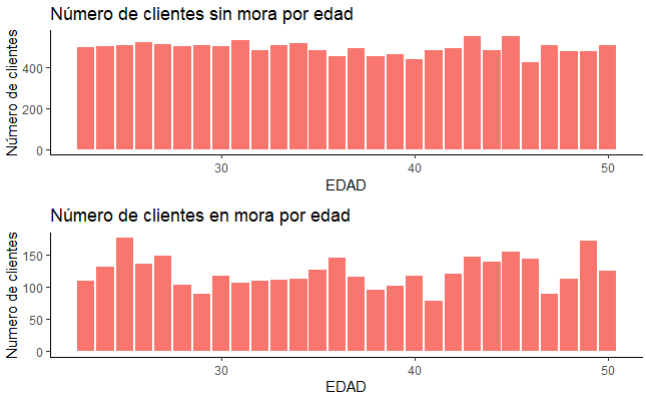


Figura 10: Número de clientes por edad

Como se puede observar en el siguiente gráfico el método del codo sugiere tomar en consideración 4 grupos:

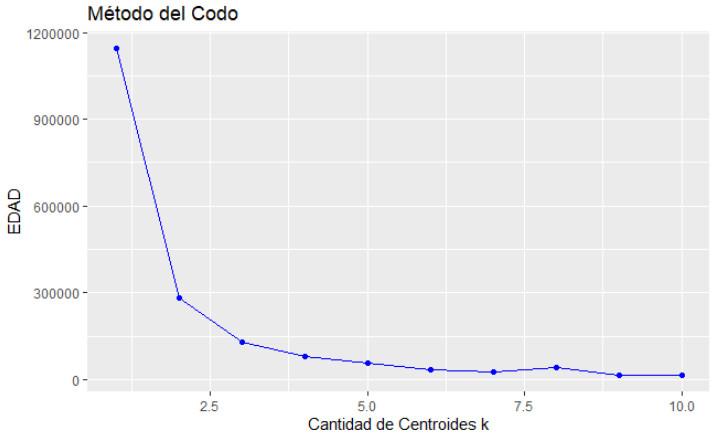


Figura 11: Categorización de Variable

Por otro lado, el método de k-medias propone tomar 3 grupos, los cuales agruparían las edades de la siguiente manera:

- Grupo A: [23,32]
- Grupo B: [33,41]
- Grupo C: [42,50]

7.2. Variables Categóricas

Se realiza una reducción de categorías para las variables **INSTRUCCIÓN, SEGMENTO_RIESGO y SUCURSAL**. Para este ejercicio se hace uso de la función *suggest_levels*(), facilitada por el software R. Esta función determina niveles que sean similares entre sí en términos de porcentajes de cada nivel de una variable categórica de dos niveles. En otras palabras, esta función calcula el porcentaje de cada nivel de $y = MarcaMora\_Tarjeta$ para cada nivel de x.

- **INSTRUCCIÓN**

ID	BIC	Cluster
1	5382.106	all in one
2	5384.190	(SEC)(UNI&TEC&PRI)
3	5389.310	(SEC)(UNI&TEC)(PRI)
4	5398.398	(SEC)(UNI)(TEC)(PRI)

Entonces se pueden agrupar las categorías de la siguiente forma:

$$clusters : (SEC)(UNI\&TEC)(PRI)$$

- **SEGMENTO_RIESGO**

ID	BIC	Cluster
1	5382.106	all in one
2	5347.990	(D&C&B&E)(A)
3	5235.407	(D&C)(B&E)(A)
4	4998.410	(D)(C)(B&E)(A)
5	5007.763	(D)(C)(B)(E)(A)

Entonces se pueden agrupar las categorías de la siguiente forma:

$$clusters : (D\&C)(B\&E)(A)$$

- **SUCURSAL**

  Para la variable sucursal, se procede de forma similar y se obtiene que la mejor forma de agruparlos sería en dos grupos:

  - Costa: Esmeraldas, Machala, Manta, Guayaquil , Quevedo.
  - Sierra: Cuenca, Ibarra, Quito, Ambato, Loja, Latacunga, Santo Domingo, Riobamba.

8. Selección de las variables explicativas

8.1. Variables numéricas

Prueba de Kolmogorov-Smirnov (KS)

El test de Kolmogorov-Smirnov es una prueba de bondad de ajuste, donde se contrasta la hipótesis de si las muestras aleatorias e independientes provienen de distribuciones continuas idénticas.

A continuación se describe esta prueba para dos muestras aleatorias donde se contrastan las hipótesis:

$$\begin{cases} Ho : F_1(x) = F_2(x) \\ \\ H_1 : F_1(x) \neq F_2(x) \end{cases}$$

En otras palabras, KS mide el grado de concordancia existente entre la distribución de un conjunto de datos y la distribución teórica específica. Su objetivo es evaluar si una muestra podría haber sido muestreada a partir de una distribución de probabilidad específica. Sea $G_n(t)$ la función de distribución acumulada empírica para la muestra y F_t la función de distribución acumulada teórica (Berger Zhou, 2014)[3]. El test toma la siguiente forma:

$$D_k = max|F(t) - G_n(t)|,\ min(x) \leq t \leq max(x)$$

En los resultados si el nivel de significancia es menor a 0,05 la distribución no es normal y si es mayor a 0,05, si lo es. En la siguiente tabla se puede observar el estadístico de KS:

Variable	KS
SALDO_TOTAL_TARJETA	0.3122
ANTIGUEDAD_TARJETA_ANIOS	0.3087
PROMEDIO_MENSUAL_CONSUMOS_LOCALES	0.1895
NUMERO_OPERACIONES_TITULAR	0.1690
PROMEDIO_DIAS_SOBREGIRO_CC	0.1485
PROMEDIO_MENSUAL_SALDO_CUENTA_PASIVO	0.1221
RIESGO_CLIENTE_TOTAL_GFP	0.1109
VALOR_DEPOSITO_A_PLAZO	0.0361
NUM_TC_SIST_FIM	0.0281
relacion_saldo_cupo	0.0257

Las distribuciones de las variables con mayor valor en el estadístico KS son aquellas con mayor potencial predictivo. Se aplica el criterio de seleccionar a las variables que al menos tomen el valor de 5 % en el KS.

8.2. Variables categóricas

Para filtrar las variables predictoras, previo a la regresión logística, se utiliza el estadístico "valor de información" (VI). El VI es una técnica útil para seleccionar variables importantes en un modelo de regresión logística binaria y para clasificarlas en función de su importancia. El VI se compone del estadístico "weight of evidence" (WOE), el cual indica el poder predictivo de una variable independiente en relación con la variable dependiente. WOE también puede ser interpretado como una medida de separación entre buenos y malos. Los clientes malos serían aquellos que incumplen con el préstamo (% of events) y los buenos los que si pagaron (%of non_events). A continuación, se puede apreciar la fórmula del WOE y del VI:

$$WOE = ln\left(\frac{\%\ of\ non\ events}{\%ofevents}\right)$$

$$VI = \sum_i (\%\ of\ non\ events_i - \%\ of\ events_i) * WOE$$

Respecto a la interpretación de los resultados del VI, se ha tomado como referencia el rango de 0,3 hasta 0,5 como un indicador de un fuerte poder predictivo (Siddiqi, 2006, como se citó en Lund Brotherton, 2013)[16].

Valor de la información	**Predictividad variable**
Menos de 0.02	No es útil para la predicción
0.02 hasta 0.1	Poder predictivo débil
0.1 hasta 0.3	Poder predictivo medio
0.3 hasta 0.5	Fuerte poder predictivo
Más de 0.5	Poder predictivo sospechoso

Tabla 3: Interpretación de resultados IV

Al aplicar el VI en las variables categóricas, se decide tomar las variables con un VI entre 0.02 y 0.5, ya que éstas cuentan con un poder predictivo. Es importante notar que la variable cantidad_total_avances lidera los resultados, ya que obtiene el mayor valor de VI. Según este método las variables menos influyentes serían Instrucción, segmento riesgo y sucursal.

Variable	**VI**
CANTIDAD_TOTAL_AVANCES	0.1539526
grupo_MÁXIMO_NUM_DIAS_VENCIDO	0.0892575
FORMA_PAGO	0.0535041
MARCA_CUENTA_CORRIENTE	0.0320038
MARCA_CUENTA_AHORROS	0.0320038
ORIGEN_APROBACIÓN	0.0077591
grupo_EDAD	0.0049004
GENERO	0.0002275
INSTRUCCIÓN	0.0000060
SEGMENTO_RIESGO	0.0000000
SUCURSAL	0.0000000

9. Regresión Logística

La Regresión Logística es una técnica estadística multivariante que nos permite estimar la relación existente entre una variable dependiente no métrica, en particular dicotómica y un conjunto de variables independientes métricas o no métricas. El Análisis de Regresión Logística tiene la misma estrategia que el Análisis de Regresión Lineal Múltiple, el cual se diferencia esencialmente del Análisis de Regresión Logística por que la variable dependiente es métrica.

La variable dependiente o respuesta no es continua, sino discreta (generalmente toma valores 1,0). Las variables explicativas pueden ser cuantitativas o cualitativas; y la ecuación del modelo no es una función lineal de partida, sino exponencial; si bien, por sencilla transformación logarítmica, puede finalmente presentarse como una función lineal. Así pues el modelo será útil en frecuentes situaciones prácticas de investigación en que la respuesta puede tomar únicamente dos valores: 1, presencia (con probabilidad p); y 0, ausencia (con probabilidad 1-p) (Salcedo, s.f.)[15].

9.1. Regresión Logística Simple

La Regresión Logística Simple, desarrollada por David Cox en 1958, es un método de regresión que permite estimar la probabilidad de una variable cualitativa binaria en función de una variable cuantitativa. Una de las principales aplicaciones de la regresión logística es la de clasificación binaria, en el que las observaciones se clasifican en un grupo u otro dependiendo del valor que tome la variable empleada como predictor.

¿Por qué regresión logística y no lineal?

Si una variable cualitativa con dos niveles se codifica como 1 y 0, matemáticamente es posible ajustar un modelo de regresión lineal por mínimos cuadrados $\beta_0 + \beta_1 X$. El problema de esta aproximación es que, al tratarse de una recta, para valores extremos del predictor, se obtienen valores de Y menores que 0 o mayores que 1, lo que entra en contradicción con el hecho de que las probabilidades siempre están dentro del rango [0,1].

la regresión logística transforma el valor devuelto por la regresión lineal ($\beta_0 + \beta_1 X$ empleando una función cuyo resultado está siempre comprendido entre 0 y 1. Existen varias funciones que cumplen esta descripción, una de las más utilizadas es la función logística (también conocida como función sigmoide):

$$función : \sigma(x) = \frac{1}{1 + e^{-x}}$$

Para valores de x muy grandes positivos, el valor de e^x es aproximadamente 0 por lo que el valor de la función sigmoide es 1. Para valores de x muy grandes negativos, el valor e^x tiende a infinito por lo que el valor de la función sigmoide es 0.

Sustituyendo la x de la ecuación anterior por la función lineal ($\beta_0 + \beta_1 X$) se obtiene que:

$$\begin{aligned} P(Y = k|X = x) &= \frac{1}{1 + e^{-(\beta_0+\beta_1 X)}} \\ &= \frac{1}{\frac{e^{\beta_0+\beta_1 X}}{e^{\beta_0+\beta_1 X}} + \frac{1}{e^{\beta_0+\beta_1 X}}} \\ &= \frac{1}{\frac{1+e^{\beta_0+\beta_1 X}}{e^{\beta_0+\beta_1 X}}} \\ &= \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \end{aligned}$$

donde $Pr(Y = k|X = x)$ puede interpretarse como: la probabilidad de que la variable cualitativa Y adquiera el valor k (el nivel de referencia, codificado como 1), dado que el predictor X tiene el valor x.

Esta función, puede ajustarse de forma sencilla con métodos de regresión lineal si se emplea su versión logarítmica, obteniendo lo que se conoce como LOG of ODDs

$$ln\left(\frac{p(Y = k|X = x)}{1 - p(Y = k|X = x)}\right) = \beta_0 + \beta_1 X$$

10. Metodología

10.1. Regresión Logística Múltiple

La regresión logística múltiple es una extensión de la regresión logística simple. Se basa en los mismos principios que la regresión logística simple (explicados anteriormente), pero ampliando el número de predictores. Los predictores pueden ser tanto continuos como categóricos.

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i$$

$$logit(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i$$

El valor de la probabilidad de Y se puede obtener con la inversa del logaritmo natural:

$$p(Y) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_i x_i}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_i x_i}}$$

A la hora de evaluar la validez y calidad de un modelo de regresión logística múltiple, se analiza tanto el modelo en su conjunto como los predictores que lo forman(Amat,2016)[1].

11. Selección de variables

Se van a seleccionar las variables explicativas que presentan mayor divergencia entre las distribuciones de buen y mal pagador, considerando los valores de KS y VI mostrados en la sección 7.

Para las variables numéricas se calculó el estadístico KS, el cual sugiere seleccionar todas las variables a excepción de las variables Valor depósito a plazo, num tc sist fim y relación saldo cupo tarjeta, ya que las mismas tienen un porcentaje menor al 5 %. No obstante, se analizará y se probará su inclusión en el modelo logístico.

Para las variables categóricas se utilizó el índice de Valor de Información (VI) y en función de este estadístico se deberían seleccionar las variables marca cuenta ahorros, marca cuenta corriente, forma pago y cantidad total de avances ya que a pesar de tener un poder predictivo débil, si lo tienen. Sin embargo, al igual que con KS, se analizará y probará la inclusión de variables.

12. Planteamiento del modelo logit

Para la modelización se tomaron las variables descritas en la sección 9, sin embargo algunas de estas variables resultaron no ser significativas entonces se procedió a retirarlas del modelo de forma ordenada.

Se han construido 2 posibles modelos, para compararlos se ha recurrido a utilizar el criterio de información de Akaike (AIC). Éste criterio es una medida de bondad de ajuste que describe la relación entre el sesgo y la varianza en la construcción del modelo. Dicho de otra forma califica la exactitud y complejidad del modelo, por lo que un mayor valor del AIC sugiere un mejor modelo. De forma análoga, se aplica también el estadístico de desviación residual. Este valor describe la desviación estándar de los puntos formados alrededor de una función lineal y es una estimación de la precisión de la variable dependiente que se mide. Por lo tanto un mayor valor de este estadístico responde a un mejor modelo.

Indicador	**Modelo 1**	**Modelo 2**
AIC	4670.8	4646.7
Residual deviance	4654.8	4628.7

En la anterior tabla se puede ver que el segundo modelo es el más adecuado, debido a que los estadísticos evaluados son mayores en el segundo modelo. Entonces, para la formación del modelo final se tiene las siguientes variables:

	Estimate	Std. Error	z value	Pr(> ‖z‖)
(Intercept)	3.463e-01	1.147e-01	3.020	0.002526
SALDO_TOTAL_TARJETA	1.351e-04	1.411e-05	9.578	<2e-16
ANTIGUEDAD_TARJETA_ANIOS	-2.302e-01	8.189e-03	-28.115	<2e-16
PROMEDIO_MENSUAL_CONSUMOS_LOCALES	-9.705e-04	1.041e-04	-9.324	<2e-16
NUMERO_OPERACIONES_TITULAR	3.608e-02	1.178e-02	3.061	0.002205
PROMEDIO_MENSUAL_SALDO_CUENTA_PASIVO	-6.240e-05	9.970e-06	-6.259	3.87e-10
RIESGO_CLIENTE_TOTAL_GFP	-3.869e-06	1.700e-06	-2.276	0.022844
VALOR_DEPOSITO_A_PLAZO	-1.203e-04	2.757e-05	-4.364	1.27e-05
relacion_saldo_cupo	4.017e-01	6.126e-02	6.557	5.47e-11
CANTIDAD_TOTAL_AVANCES1	3.416e-01	4.181e-02	8.170	3.07e-16
grupo_MAXIMO_NUM_DIAS_VENCIDO2	-8.346e-01	9.343e-02	-8.933	<2e-16
grupo_MAXIMO_NUM_DIAS_VENCIDO3	-4.339e-01	9.597e-02	-4.521	6.16e-06
grupo_MAXIMO_NUM_DIAS_VENCIDO4	-3.691e-01	1.007e-01	-3.665	0.000248
FORMA_PAGOSi	-4.204e-01	5.259e-02	-7.994	1.31e-15
MARCA_CUENTA_CORRIENTESi	-1.777e-01	4.237e-02	-4.194	2.75e-05
grupo_EDADG_1	-1.465e-01	5.728e-02	-2.558	0.010515
grupo_EDADG_2	-1.751e-01	5.533e-02	-3.164	0.001554
grupo_EDADG_3	-2.315e-01	6.118e-02	-3.784	0.000154
SEGMENTO_RIESGOB_E	1.640e+00	5.442e-02	30.137	<2e-16
SEGMENTO_RIESGOD_C	1.947e+00	9.424e-02	20.661	<2e-16
SUCURSALSierra	-4.987e-01	3.942e-02	-12.653	<2e-16

Los resultados evidencian que todos los coeficientes de las variables del modelo son significativos, ya que su p-valor es inferior al nivel crítico de 0,05. El signo esperado de los coeficientes casi siempre coincide con los resultados obtenidos. Por ejemplo variables como Saldo_total_tarjeta, Numero_operaciones_titular y relacion_saldo_cupo, llevan un signo positivo. Lo que quiere decir que el modelo sugiere que a mayor saldo de tarjeta, mayor número de operaciones y mayor utilización de la tarjeta en relación al cupo permitido; hay mayor probabilidad de que el cliente sea un mal pagador. Por otro lado, variables como Promedio_mensual_consumos_locales y Antiguedad_tarjeta_anios tienen el coeficiente negativo. Este resultado también resulta coherente dado que sugiere que a mientras mayor sea la antiguedad y el consumo del cliente; menor será la probabilidad de que sea un mal pagador. Un hallazgo que llamó nuestra atención es el coeficiente negativo de la Sucursal_Sierra, lo cual sugiere que los clientes de las ciudades registradas de esta región del país, no tienden a ser malos pagadores.

12.1. Interpretación Odds

El odd es la probabilidad de que suceda un evento dividido por la probabilidad de que no suceda. El odd ratio (razón de probabilidades) es una medida de asociación entre dos variables que indica la fortaleza de la relación entre dos variables. Los odd ratio oscilan entre 0 e infinito. Cuando el odd ratio es 1 indica ausencia de asociación entre las variables. Los valores menores a 1 señalan asociación negativa y los mayores a 1, asociación positiva. En modelos de regresión logística los odd ratios se usan para comparar la influencia de las variables explicativas sobre la variable independiente. Con el cálculo de los coeficientes b se puede saber si las variables independientes están relacionadas con la variable dependiente. Los coeficientes b se pueden expresar mediante los odd ratios, elevando el coeficiente b al número exponencial e (e^b).

La regresión logística recurre a los odd ratios porque son medidas estandarizadas que permiten comparar el nivel de influencia o fortaleza de las variables independientes sobre la variable dependiente. Las variables independientes están en diferentes escalas y necesitan ser estandarizadas para poder compararlas, ésta estandarización es el odd ratio o exponencial de b. Cuando el odd ratio es mayor a 1 un aumento de la variable independiente aumenta los odds de que ocurra el evento. Cuando es menor a 1, un aumento de la variable independiente, reduce los odds de que ocurra el evento (Cárdenas, 2015)[5]. A continuación se presenta una tabla con los odds del modelo:

(Intercept)	SALDO_TOTAL_TARJETA
1.4137917	1.0001351
ANTIGUEDAD_TARJETA_ANIOS	PROMEDIO_MENSUAL_CONSUMOS_LOCALES
0.7943465	0.9990300
NUMERO_OPERACIONES_TITULAR	PROMEDIO_MENSUAL_SALDO_CUENTA_PASIVO
1.0367339	0.9999376
RIESGO_CLIENTE_TOTAL_GFP	VALOR_DEPOSITO_A_PLAZO
0.9999961	0.9998797
relacion_saldo_cupo	CANTIDAD_TOTAL_AVANCES1
1.4943596	1.4071638
grupo_MAXIMO_NUM_DIAS_VENCIDO2	grupo_MAXIMO_NUM_DIAS_VENCIDO3
0.4340377	0.6479910
grupo_MAXIMO_NUM_DIAS_VENCIDO4	FORMA_PAGOSi
0.6913688	0.6568114
MARCA_CUENTA_CORRIENTESi	grupo_EDADG_1
0.8372084	0.8636956
grupo_EDADG_2	grupo_EDADG_3
0.8393972	0.7933401
SEGMENTO_RIESGOB__E	SEGMENTO_RIESGOD__C
5.1562359	7.0075106
SUCURSALSierra	
0.6073060	

Las variables con un odd ratio mayor a 1, donde se aumenta la probabilidad de que ocurra el evento (marca mora tarjeta), son: número operaciones titular, relación saldo cupo, segmento riesgo grupo BE, saldo total tarjeta, cantidad total de avances 1 y segmento riesgo grupo DC. El mayor odd ratio obtenido corresponde a la variable segmento riesgo grupo DC. Su interpretación es que el aumento de una unidad en el segmento riesgo, y si el resto de variables se mantuvieran constantes, aumentaría las probabilidades (los odds) de que el cliente sea un mal pagador en 5,16 veces más, que si no se aumentara en esa unidad del segmento riesgo grupo DC.

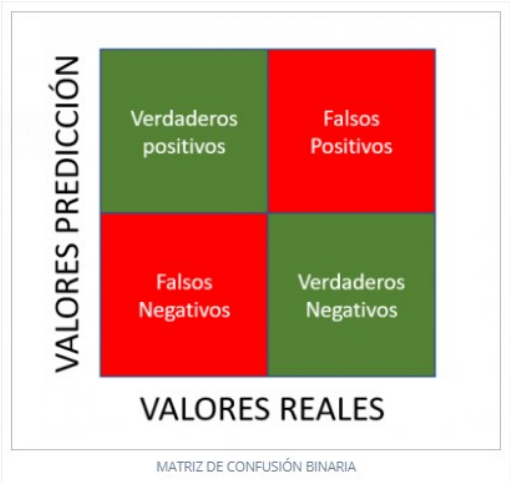
Las variables con un odd ratio menor a 1, donde se reduce la probabilidad de que ocurra el evento (marca mora tarjeta), son: antigüedad tarjeta años, riesgo cliente total gfp, máximo num días vencido grupo 2-3-4, marca cuenta corrientei, edad grupo 1-2-3, sucursal Sierra, promedio mensual consumos locales, promedio mensual saldo cuenta pasivo, valor deposito a plazo y forma pagosi.

13. Validación del modelo

Como se muestra en la tabla anterior, tenemos que todas las variables son significativas vamos analizar la matriz de confusión.

13.1. Matriz de confusión

Para validar el modelo se realiza una matriz de confusión para visualizar el desempeño predictivo de la submuestra train.



		Reference	
		0	**1**
Prediction	**0**	10.956	2.996
	1	1.019	2.252

		Reference %	
		0	**1**
Prediction	**0**	63,61 %	17,40 %
	1	5,92 %	13,08 %

Se tiene que la precisión del modelo es del 79,69 % lo cual nos dice que el modelo está clasificando correctamente.

13.2. Indicadores

13.2.1. Kolmogorov-Smirnov (KS)

La prueba KS ayuda a medir el poder predictivo de un modelo de clasificación[3].En teoría el KS debería tomar valores entre 0.2 y 0.7 frente a una buena capacidad de clasificación del modelo. Por debajo de 0.2 el modelo resulta con bajo poder predictivo y por encima de 0.7 el poder predictivo sería sospechosamente muy bueno (Yépez, 2019)[18].

Además, el valor del estadístico K-S debe ser mayor al 20 % para obtener un buen ajuste de los datos, del modelo mencionado se obtuvo un $KS = 0{,}4727$ Por lo tanto, el modelo propuesto genera una alta confiabilidad sobre la predicción de los datos.

13.2.2. Curva de ROC

La curva ROC, o curva característica operativa del recepto, es una representación gráfica del rendimiento de un modelo de clasificación en todos los umbrales de discriminación. La curva representa dos parámetros, la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR).

$$TPR = \frac{Verdaderos\ positivos}{(Verdaderos\ positivos\ +\ Falsos\ negativos)}$$

[3]estadístico abordado en la sección 7.1

$$FPR = \frac{Falsos\ positivos}{(Falsos\ positivos\ +\ Verdaderos\ negativos)}$$

ROC se construye a partir de las submuestras "train" y "test" y el área bajo esta curva (AUC) es un indicador de discriminación que proporciona una medición del rendimiento en todos los umbrales de clasificación. AUC es la probabilidad de que el modelo clasifique un ejemplo positivo aleatorio más alto que un ejemplo negativo aleatorio. En este sentido, AUC toma valores entre 0 y 1, 0 si las predicciones son en un 100 % incorrectas y 1 si las predicciones son 100 % correctas (Google IA, 2020)[12].

En el presente modelo se obtuvo la siguiente curva ROC:

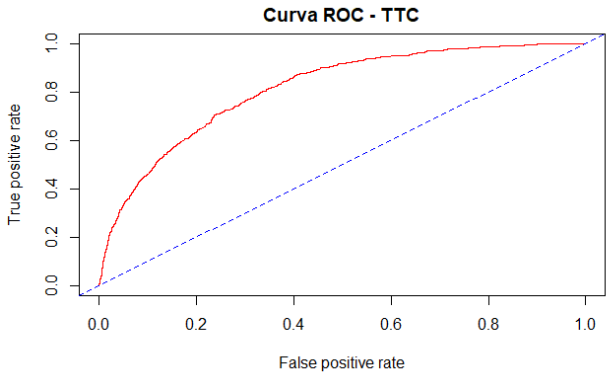


Figura 12: curva ROC - TTC

Podemos observar que la calidad de discriminación del modelo es buena ya que la curva difiere en gran medida de la recta de clasificación aleatoria.

13.2.3. Curva de Lorenz y Coeficiente de Gini

La curva de Lorenz grafica la fracción acumulada de una variable aleatoria versus la fracción acumulada de la población receptora de esa variable repartida. Cuanto más alejada se encuentre la curva de Lorenz de la línea de igualdad perfecta, mayor es la desigualdad que se presenta. El coeficiente de Gini se deriva de la curva de Lorenz y es una medida de desigualdad en la distribución de una variable. Gini puede tomar valores entre 0 y 1, donde 0 representa igualdad perfecta y uno desigualdad perfecta.

Calculando el Gini de este modelo, se obtiene un resultado de 81,11 % para la base de entrenamiento, por lo que se puede decir que el ajuste del modelo es bueno como se observó en la gráfica de ROC.

14. Factor de inflación de la varianza (FIV)

Los factores de inflación de la varianza (FIV) son usados para comprobar que no existe colinealidad/multicolinealidad entre las variables predictoras. El FIV proporciona un índice que mide hasta qué punto la varianza de un coeficiente de regresión estimado se incrementa a causa de la colinealidad. La función se define de la siguiente manera:

$$VIF = \frac{1}{1 - R^2}$$

Donde el R^2 es el estadístico que cuantifica el grado en que un predictor se correlaciona con las otras variables predictoras en una regresión lineal. Un valor de 1 significa que el predictor no está correlacionado con otras

variables. Mientras mayor sea el valor, mayor es la correlación de la variable con otras variables. Valores mayores a 4 son considerados moderados a altos y los valores mayores a 10, muy altos. Cuanto mayor sea el FIV, más se inflará el error estándar, mayor será el intervalo de confianza y menor la probabilidad de que se determine que un coeficiente es estadísticamente significativo. Lo que significa que sería difícil o imposible evaluar con precisión la contribución de los predictores a un modelo (Bock, s.f)[4]. En la siguiente tabla se puede observar que el FIV para las variables numéricas, se encuentra siempre alrededor de 1, por lo que se puede concluir que no hay multicolinealidad.

Variables Numéricas	**VIF**
SALDO_TOTAL_TARJETA	1.250425
ANTIGUEDAD_TARJETA_ANIOS	1.285474
PROMEDIO_MENSUAL_CONSUMOS_LOCALES	1.138035
NUMERO_OPERACIONES_TITULAR	1.187407
PROMEDIO_MENSUAL_SALDO_CUENTA_PASIVO	1.070879
RIESGO_CLIENTE_TOTAL_GFP	1.123627
VALOR_DEPOSITO_A_PLAZO	1.005104
relacion_saldo_cupo	1.318772

El FIV puede ser aplicado únicamente en variables numéricas; sin embargo, una versión generalizada del FIV (FIVG), usada para probar conjuntos de variables predictoras y modelos lineales generalizados; es útil para variables categóricas. En la siguiente tabla se puede observar que el FIV generalizado para las variables categóricas está alrededor de 1. Dado que es menor a 4, no se presenta multicolinealidad.

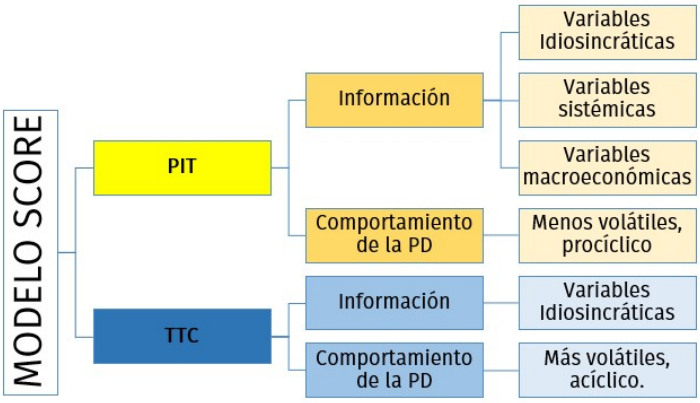
Variable Categóricas	**GVIF**
CANTIDAD_TOTAL_AVANCES	1.044010
grupo_MAXIMO_NUM_DIAS_VENCIDO	1.045175
FORMA_PAGO	1.067449
MARCA_CUENTA_CORRIENTE	1.099530
Grupo_EDAD	1.002367
SEGMENTO_RIESGO	1.023133
SUCURSAL	1.011350

15. Filosofías para el modelo Logit

Para estimar la PD (Probability of default[4]) de un cliente, las empresas financieras buscan generar un modelo que les permita clasificar a los clientes en grupos de riesgo, de acuerdo a la información que la institución disponga sobre cada uno de ellos. Los sistemas de clasificación crediticia asignan a los clientes en diferentes grupos de riesgo, con el fin de distinguirlos en función de su calidad crediticia.

Los sistemas de clasificación pueden tener un enfoque PIT (Point in time) o TTC (Through the cycle) y la diferencia entre ambos está en la información que el score de cada sistema usa; tal como se lo puede apreciar en el siguiente gráfico:

[4]La PD es una medida de calificación crediticia que se otorga internamente a un cliente o a un contrato con el objetivo de estimar su probabilidad de incumplimiento a un año vista.



15.1. Filosofía Through the cycle *TTC*

No existe una definición concisa de TTC, pero generalmente hacen referencia a las estimaciones de PD estresadas. De hecho, se manifiesta que una PD estresada se destaca por tener un enfoque TTC, mientras que una PD no estresada pertenece al enfoque PIT (Bassel Commmite on Banking Supervision, 2005).

Una medida de riesgo TTC, a diferencia de la PIT, es menos precisa pero tiene un alto grado de estabilidad y fluidez. Esta estabilidad se produce a costa de una menor puntualidad y precisión de predicción predeterminada en relación con las medidas de riesgo de PIT. Los PD de TTC resultan valiosos cuando los costos de ajuste de la cartera o los costos de cumplimiento normativo son altos (Hamilton, 2011)[9]. El tipo de información que este enfoque utiliza es principalmente idiosincrático. La PD atribuida al individuo a lo largo del tiempo no acompaña necesariamente al ciclo de crédito y al ciclo económico. Los individuos a lo largo del tiempo, tienden a permanecer estables. La tasa de morosidad observada del grupo, tiende a ser más volátil que la de PIT.

15.1.1. Creación de los grupos de riesgo (Clustering)

Para la creación de los grupos se aplicará un análisis tipo clúster. Clustering es el proceso de dividir un conjunto de objetos de datos (u observaciones) en subconjuntos. Cada subconjunto es un clúster, de modo que los objetos de un clúster son similares entre sí, pero diferentes a los objetos de otros clústeres. El uso de este algoritmo de agrupación permite descubrir grupos previamente desconocidos dentro de los datos (Hiawei et al, 2012)[10].

15.1.2. Calidad del Cluster

Para evaluar la calidad de los clúster se aplica el método del codo, el cual se lo puede observar plasmado en la siguiente gráfica:

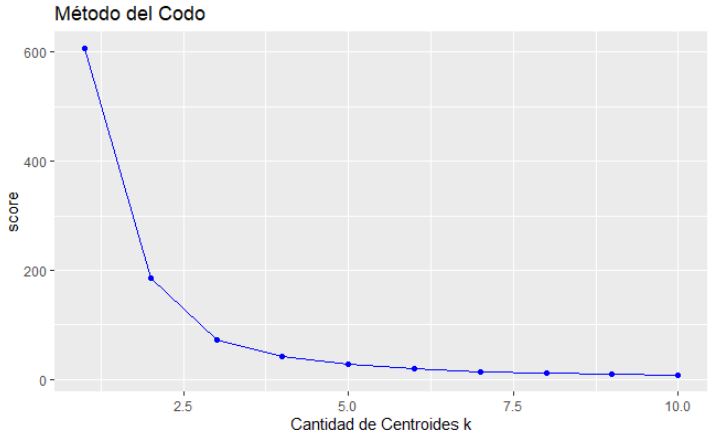


Figura 13: Elección del numero de grupos

Es así que se plantea considerar 3 grupos, los cuales fueron hallados por k-medias. Y se distribuyen de la siguiente manera:

- GH 1: [0; 0.2293]
- GH 3: [0.2295; 0.5405]
- GH 3: [0.5411; 0.9950)

A continuación se hace un box plot de los grupos, para evidenciar que los grupos posean diferente media y estén separados entre sí.

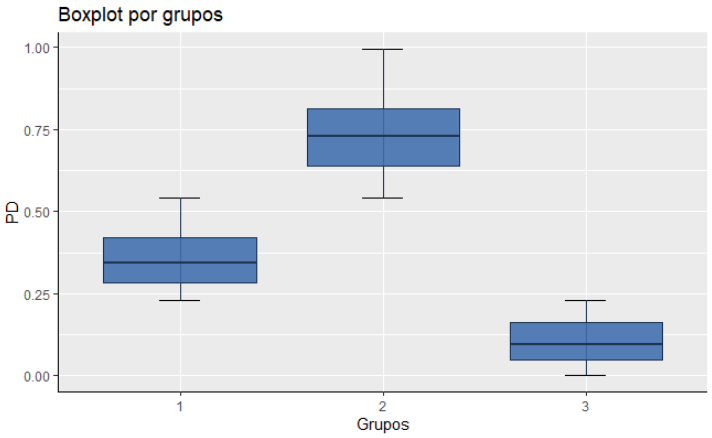


Figura 14: Diagrama de caja de los grupos

A continuación se muestra un resumen de como están distribuidos los grupos de riesgo:

Grupo	GH 1
0	1
774	1.185

Grupo	GH 2
0	1
8.398	511

Grupo	GH 1
0	1
4.617	1.738

Además, se asigna una PD para cada grupo que se ha encontrado:

GH	PD
1	0.3406306
2	0.9061695
3	0.6242302

15.1.3. Validación de los clústers

Con el fin de evaluar la calidad de agrupamiento de los clústers propuestos en el análisis anterior se emplearán tres estadísticos, Davies-Bouldin, Dunnett T3, test de Bartlett, test Games y Howell.

Índice Davies-Bouldin

El criterio de Davies-Bouldin se basa en la relación entre las distancias "dentro del grupoz .entre grupos":

$$DB(C) = \frac{1}{k}\sum_{i=1}^{k} \underset{j\leq k, j\neq i}{max} D_{ij}, k = |C|\,,$$

D_{ij} es el ratio de distancia de los cluster dentro y entre, para los grupos i y j. D_i es la distancia promedio entre cada punto de datos en el grupo i y su centroide. d_{ij} es la distancia euclidiana entre los centroides de los dos grupos.

$$D_{ij} = \frac{(\bar{d}_i + \bar{d}_j)}{d_{ij}},$$

En el caso de que dos conglomerados tuvieran una pequeña distancia y una gran dispersión, éstos no serían muy distintos. En otras palabras, un índice de Davies-Bouldin pequeño, sugiere una una óptima agrupación (Drakos, 2020)[8].

DB	0,479649

El valor del índice es 0,4911241, por lo que se sospecha que nuestros cluster están clasificando correctamente a los datos.

Prueba de Dunnett T3

El método de Dunnet compara las medias de varios grupos experimentales con la media de un grupo de control para evaluar sus diferencias. La prueba de Dunnet T3 realiza una comparación por parejas, basada en el módulo máximo estudentizado (studentized). Esta prueba resulta apropiada con tamaños de muestra pequeños (n<50). La prueba se define así:

$$\frac{(\bar{Y}_i - \bar{Y}_j)^2}{\frac{S_i^2}{n_i} + \frac{S_j^2}{n_j}} \sim t - Student$$

Donde, $\bar{Y}_i$ y $\bar{Y}_j$ son las medias muestrales del i-ésimo y j-ésimo grupo, S_i^2 y S_j^2 son las varianzas muestrales del i-ésimo y j-ésimo grupo y n_i y n_j son los tamaños de muestra respectivos de los grupos i y j. Esta prueba asume normalidad y homocedasticidad y su hipótesis nula es que las medias de los grupos son iguales. Por lo que un p-valor < 0.05 rechazaría H_0 y permitiría concluir que las medias de los grupos son diferentes (Yépez, 2019)[18].
Después de implementar el test Dunnet T3, se obtuvo un p-valor de 2e-16; es decir, en cada instante de tiempo

las medias de los grupos son diferentes estadísticamente, por lo tanto se puede concluir que los 3 grupos son heterogéneos entre ellos.

El **test de Bartlett** permite contrastar la igualdad de varianza en 2 o más poblaciones sin necesidad de que el tamaño de los grupos sea el mismo. Es más sensible que el test de Levene a la falta de normalidad, pero si se está seguro de que los datos provienen de una distribución normal, es la mejor opción.

Bartlett's	**K-squared**	**df**	**p-value**
GH1 - GH2	**246.25**	**1**	**2.2e-16**
GH1 - GH3	**336.38**	**1**	**2.2e-16**
GH2 - GH3	**969.98**	**1**	**2.2e-16**

Dado que el p_valor es menor a un $\alpha = 0{,}05$, se rechaza la hipótesis nula y, en consecuencia, las varianzas entre los grupos son diferentes.

El **test de Games y Howell** tiene un enfoque no paramétrico para comparar combinaciones de grupos o tratamientos. Aunque es bastante similar a la prueba de Tukey en su formulación, la prueba de Games-Howell no asume varianzas y tamaños de muestra iguales.

group1	**group2**	**estimate**	**conf.low**	**conf.high**	**p.adj**	**p.adj.signif**
GH1	GH2	0.3727301	0.3675020	0.3779583	4.19e-08	****
GH1	GH3	-0.2503821	-0.2533518	-0.2474124	0.00e+00	****
GH2	GH3	-0.6231122	-0.6281045	-0.6181199	9.27e-09	****

En consecuencia, se ve que los grupos homogéneos son heterogéneos entre sí.

Finalmente, se calcula la tasa de mora de cada grupo y se ve su comportamiento a través del tiempo del modelo PIT.

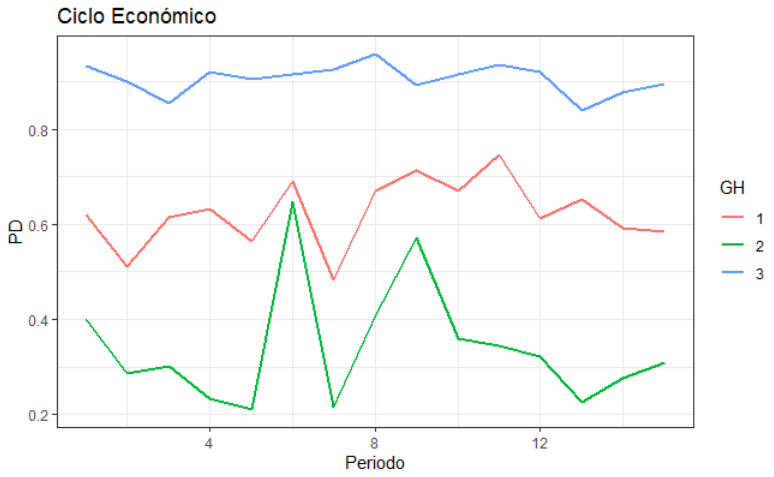


Figura 15: Tasa de mora en el tiempo de los GH

Para evidenciar si existe migración de grupos se realiza la siguiente gráfica:

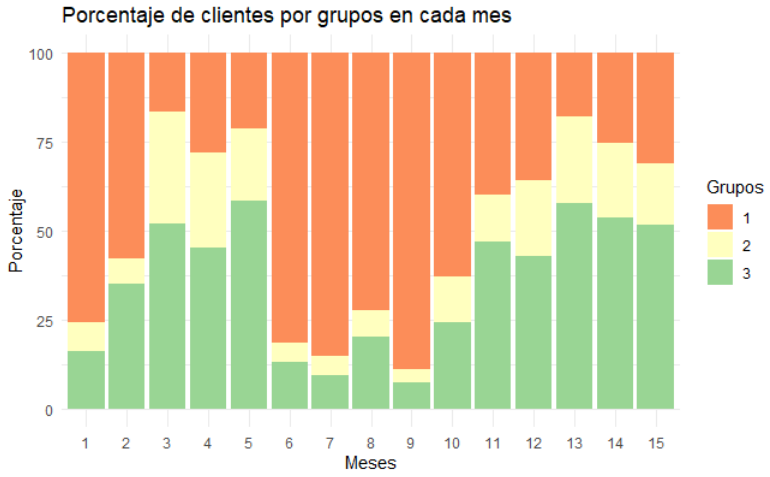


Figura 16: Porcentaje de clientes por grupo

En la gráfica precedente no se evidencia una importante migración de los individuos mes a mes. Este resultado va acorde con la teoría de la filosofía TTC, ya que dado que en este enfoque se utilizan principalmente variables idiosincráticas, es más difícil que cambien estos aspectos en los individuos a lo largo del tiempo; por este motivo se espera que los individuos migren menos.

15.2. Filosofía Point in time *PIT*

Una medida de riesgo PIT es aquella que utiliza toda la información disponible y pertinente a una fecha determinada, para estimar la probabilidad de incumplimiento esperada de una empresa. Los PD de la filosofía PIT son ideales para situaciones en las que el costo de los incumplimientos o los cambios en el diferencial del crédito es alto, por lo que es importante la detección temprana de cambios en el riesgo de crédito (Hamilton, 2011)[9]. Las calificaciones PIT intentan evaluar la situación actual de un cliente teniendo en cuenta los efectos tanto cíclicos como permanentes.

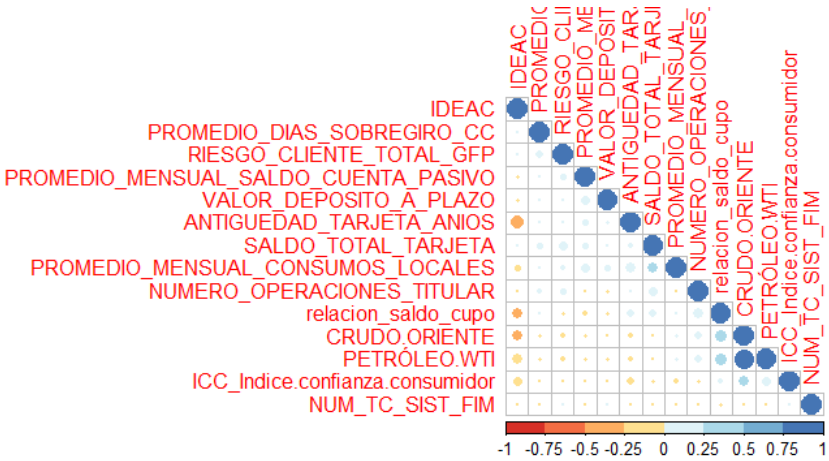


Figura 17: Correlación entre las variables

Se puede observar una alta correlación entre las variables: $PETROLEO.WTI$ y $CRUDO.ORIENTE$, lo cual puede generar problemas de multicolinealidad en el modelo logit que se va a encontrar.

Estadístico KS

En este modelo el estadístico KS sugiere tomar en cuenta todas las variables numéricas, menos valor depósito a plazo, num tc sist fim y relación saldo cupo; ya que su valor es menor a 0,05. Sin embargo, al igual que en el otro enfoque, se analizará y probará su inclusión en el modelo logístico.

Variable	KS
ICC_Indice.confianza.consumidor	0.3122
IDEAC	0.3087
CRUDO.ORIENTE	0.1895
PETRÓLEO.WTI	0.1749
SALDO_TOTAL_TARJETA	0.1566
ANTIGUEDAD_TARJETA_ANIOS	0.1483
PROMEDIO_MENSUAL_CONSUMOS_LOCALES	0.1440
NUMERO_OPERACIONES_TITULAR	0.1440
PROMEDIO_DIAS_SOBREGIRO_CC	0.1196
PROMEDIO_MENSUAL_SALDO_CUENTA_PASIVO	0.1141
RIESGO_CLIENTE_TOTAL_GFP	0.0908
VALOR_DEPOSITO_A_PLAZO	0.0375
NUM_TC_SIST_FIM	0.0348
relacion_saldo_cupo	0.0257

Valor de información (VI)

El valor de información de este modelo sugiere tomar aquellas que tienen un VI entre 0,02 y 0,5 ya que sólo éstas contarían con poder predictivo. Dicho de otra forma, calificarían las variables marca cuenta ahorros, marca cuenta corriente, forma pago, máximo num días vencido y cantidad total avances.

Variable	VI
CANTIDAD_TOTAL_AVANCES	0.1539457
grupo_MAXIMO_NUM_DIAS_VENCIDO	0.0992932
FORMA_PAGO	0.0535410
MARCA_CUENTA_CORRIENTE	0.0333463
MARCA_CUENTA_AHORROS	0.0333463
ORIGEN_APROBACION	0.0109418
grupo_EDAD	0.0035077
GENERO	0.0002275
INSTRUCCION	0.0000000
SEGMENTO_RIESGO	0.0000000
SUCURSAL	0.0000000

El tipo de información que este enfoque utiliza es idiosincrático y sistémico. La PD atribuida a lo largo del tiempo tiende a subir en periodos de recesión y a caer en periodos de expansión. Los individuos idiosincráticamente diferentes pueden tener eventualmente la misma clasificación. La tasa de morosidad observada del grupo a lo largo del tiempo, tiende a ser menos volátil.

Para la clasificación de este modelo se han tomado en consideración tanto las variables idiosincráticas como sistémicas (macroeconómicas); ICC (Índice de confianza al consumidor), IDEAC, Crudo Oriente y Petróleo WTI. La siguiente tabla expone la estimación del modelo:

	Estimate	**Std. Error**	**z value**	$Pr(>\|z\|)$
(Intercept)	-3.379e-02	6.262e-01	-0.054	0.956974
IDEAC	9.965e-03	2.987e-03	3.336	0.000849
CRUDO.ORIENTE	-2.341e-02	3.655e-03	-6.404	1.51e-10
SALDO_TOTAL_TARJETA	1.278e-04	1.417e-05	9.014	<2e-16
ANTIGUEDAD_TARJETA_ANIOS	-2.310e-01	9.200e-03	-25.108	<2e-16
PROMEDIO_MENSUAL_CONSUMOS_LOCALES	-8.795e-04	1.031e-04	-8.532	<2e-16
NUMERO_OPERACIONES_TITULAR	3.986e-02	1.185e-02	3.365	0.000766
PROMEDIO_MENSUAL_SALDO_CUENTA_PASIVO	-5.980e-05	9.864e-06	-6.062	1.34e-09
RIESGO_CLIENTE_TOTAL_GFP	-4.763e-06	1.886e-06	-2.525	0.011567
VALOR_DEPOSITO_A_PLAZO	-1.214e-04	2.748e-05	-4.417	1.00e-05
relacion_saldo_cupo	6.769e-01	6.965e-02	9.718	<2e-16
CANTIDAD_TOTAL_AVANCES1	3.184e-01	4.198e-02	7.584	3.36e-14
grupo_MAXIMO_NUM_DIAS_VENCIDO2	-9.091e-01	9.411e-02	-9.660	<2e-16
grupo_MAXIMO_NUM_DIAS_VENCIDO3	-4.319e-01	9.584e-02	-4.507	6.59e-06
grupo_MAXIMO_NUM_DIAS_VENCIDO4	-3.529e-01	1.005e-01	-3.511	0.000446
FORMA_PAGOSi	-4.185e-01	5.412e-02	-7.732	1.06e-14
MARCA_CUENTA_CORRIENTESi	-2.140e-01	4.320e-02	-4.953	7.31e-07
ORIGEN_APROBACIONProactivo	1.311e-01	5.093e-02	2.574	0.010061
grupo_EDADG_1	-1.413e-01	5.743e-02	-2.460	0.013886
grupo_EDADG_2	-1.682e-01	5.551e-02	-3.030	0.002446
grupo_EDADG_3	-2.209e-01	6.132e-02	-3.603	0.000315
SEGMENTO_RIESGOB_E	1.596e+00	5.466e-02	29.195	<2e-16
SEGMENTO_RIESGOD_C	1.904e+00	9.448e-02	20.150	<2e-16
SUCURSALSierra	-4.813e-01	3.962e-02	-12.146	<2e-16

Los resultados evidencian que todos los coeficientes de las variables del modelo son significativos, ya que su p-valor es inferior al nivel crítico de 0,05. El signo esperado de los coeficientes casi siempre coincide con los resultados obtenidos. Por ejemplo variables como Saldo total tarjeta, Número operaciones titular y relación saldo cupo, llevan un signo positivo. Lo que quiere decir que el modelo sugiere que a mayor saldo de tarjeta, mayor número de operaciones y mayor utilización de la tarjeta en relación al cupo permitido; hay mayor probabilidad de que el cliente sea un mal pagador. Cabe denotar que en este grupo de resultados con signos positivos se encuentra también la variable sistémica IDEAC, la cual es una proyección del PIB. Este resultado sugiere que a mayor volumen de actividad económica del país, mayor posibilidad de que el cliente sea un mal pagador.

Por otro lado, variables como Promedio mensual consumos locales y antigüedad tarjeta años tienen el coeficiente negativo. Este resultado también resulta coherente dado que sugiere que mientras mayor sea la antigüedad y el consumo del cliente; menor será la probabilidad de que sea un mal pagador. Un hallazgo que llamó nuestra atención es el coeficiente negativo de la Sucursal Sierra, lo cual sugiere que los clientes de las ciudades registradas de esta región del país, no tienden a ser malos pagadores.

Odds ratios

(Intercept)	IDEAC
0.9667779	1.0100148
CRUDO.ORIENTE	**SALDO_TOTAL_TARJETA**
0.9768647	1.0001278
ANTIGUEDAD_TARJETA_ANIOS	**PROMEDIO_MENSUAL_CONSUMOS_LOCALES**
0.7937425	0.9991209
NUMERO_OPERACIONES_TITULAR	**PROMEDIO_MENSUAL_SALDO_CUENTA_PASIVO**
1.0406618	0.9999402
RIESGO_CLIENTE_TOTAL_GFP	**VALOR_DEPOSITO_A_PLAZO**
0.9999952	0.9998786
relacion_saldo_cupo	**CANTIDAD_TOTAL_AVANCES1**
1.9677991	1.3749293
grupo_MAXIMO_NUM_DIAS_VENCIDO2	**grupo_MAXIMO_NUM_DIAS_VENCIDO3**
0.4029068	0.6492673
grupo_MAXIMO_NUM_DIAS_VENCIDO4	**FORMA_PAGOSi**
0.7026260	0.6580576
MARCA_CUENTA_CORRIENTESi	**ORIGEN_APROBACIONProactivo**
0.8073753	1.1400692
grupo_EDADG_1	**grupo_EDADG_2**
0.8682299	0.8451981
grupo_EDADG_3	**SEGMENTO_RIESGOB__E**
0.8017699	4.9316264
SEGMENTO_RIESGOD__C	**SUCURSALSierra**
6.7113860	0.6180095

Las variables con un odd ratio mayor a 1, donde se aumenta la probabilidad de que ocurra el evento (marca mora tarjeta), son: número operaciones titular, relación saldo cupo, segmento riesgo grupo DC y BE, IDEAC, saldo total tarjeta, cantidad total de avances 1 y origen aprobación proactivo. El mayor odd ratio obtenido corresponde a la variable segmento riesgo grupo DC. Su interpretación es que el aumento de una unidad en el segmento riesgo, y si el resto de variables se mantuvieran constantes, aumentaría las probabilidades (los odds) de que el cliente sea un mal pagador en 6,71 veces más, que si no se aumentara en esa unidad del segmento riesgo grupo DC.

Las variables con un odd ratio menor a 1, donde se reduce la probabilidad de que ocurra el evento (marca mora tarjeta), son: crudo oriente, antigüedad tarjeta años, riesgo cliente total gfp, máximo num días vencido grupo 2-3-4, marca cuenta corriente i, edad grupo 1-2-3, promedio mensual consumos locales, promedio mensual saldo cuenta pasivo, valor deposito a plazo y forma pagosi y sucursal Sierra.

Matriz de confusión

		Reference	
		0	**1**
Prediction	**0**	10.907	2.910
	1	1.068	2.338

		Reference %	
		0	**1**
Prediction	**0**	63,33 %	16,90 %
	1	6,20 %	13,57 %

Se tiene que la precisión del modelo es de 79,60 lo cual nos dice que el modelo está clasificando correctamente.

Para modelo se obtuvo un KS = 0,4677 Por lo tanto, el modelo propuesto genera una alta confiabilidad sobre la predicción de los datos.

Curva ROC

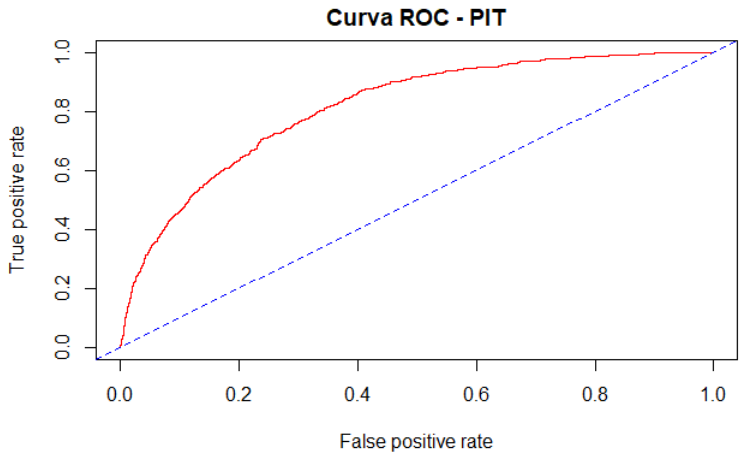


Figura 18: curva ROC - PIT

El coeficiente de Gini de este modelo, es de 81,29 % por lo que se puede decir que el ajuste del modelo es bueno, así como también se observó en la gráfica de ROC.

VIF y VIFG

Para el VIF estándar se obtuvo lo siguiente:

Variables Numéricas	VIF
IDEAC	1.662951
CRUDO.ORIENTE	1.783801
SALDO_TOTAL_TARJETA	1.263044
ANTIGUEDAD_TARJETA_ANIOS	1.668749
PROMEDIO_MENSUAL_CONSUMOS_LOCALES	1.155173
NUMERO_OPERACIONES_TITULAR	1.193463
PROMEDIO_MENSUAL_SALDO_CUENTA_PASIVO	1.073056
RIESGO_CLIENTE_TOTAL_GFP	1.142249
VALOR_DEPOSITO_A_PLAZO	1.006184
relacion_saldo_cupo	1.713095

En la tabla precedente se puede observar que el FIV para las variables numéricas, se encuentra siempre alrededor de 1, por lo que se puede concluir que no hay multicolinealidad. En cambio, para el VIF generalizado los resultados son:

Variables Numéricas	GVIF
CANTIDAD_TOTAL_AVANCES	1.046157
grupo_MAXIMO_NUM_DIAS_VENCIDO	1.057126
FORMA_PAGO	1.093122
MARCA_CUENTA_CORRIENTE	1.117852
ORIGEN_APROBACION	1.061646
grupo_EDAD	1.002563
SEGMENTO_RIESGO	1.025746
SUCURSAL	1.013840

De la misma manera que el VIF estándar los resultados del VIF generalizado se sitúan alrededor de 1 y dado que son menores a 4, se concluye que no hay presencia de multicolinealidad.

15.2.1. Creación de grupos de riesgo (Clustering)

De igual manera que en el TTC, se realizará la creación de los grupos de riesgo según el análisis clúster. En este sentido, el número óptimo de grupos bajo el sistema PIT es:

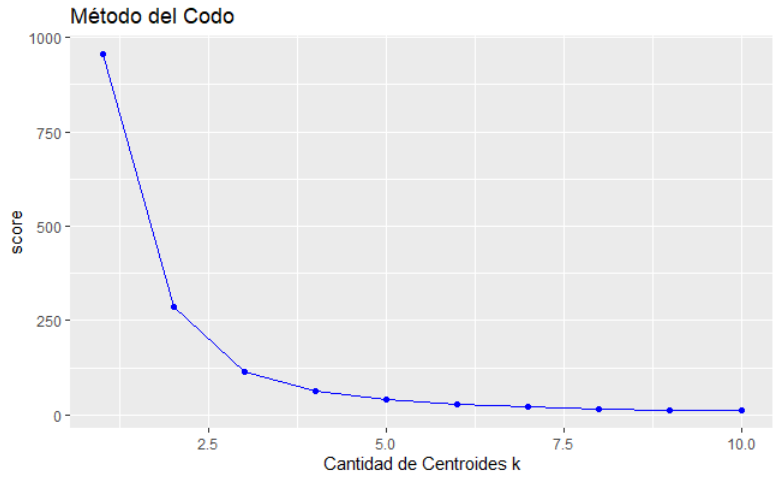


Figura 19: Elección del numero de grupos PIT

Por el método del codo, mostrado en la gráfica anterior, se deciden considerar 3 grupos, los cuales se han encontrado por k-medias y son los siguientes:

- GH1: [0.0000; 0.2331]
- GH2: [0.2332; 0.5437]
- GH3: [0.5439; 0.9995]

Con el fin de determinar la diferencia de la media y la dispersión entre los grupos, a continuación se realiza el diagrama de cajas y bigotes:

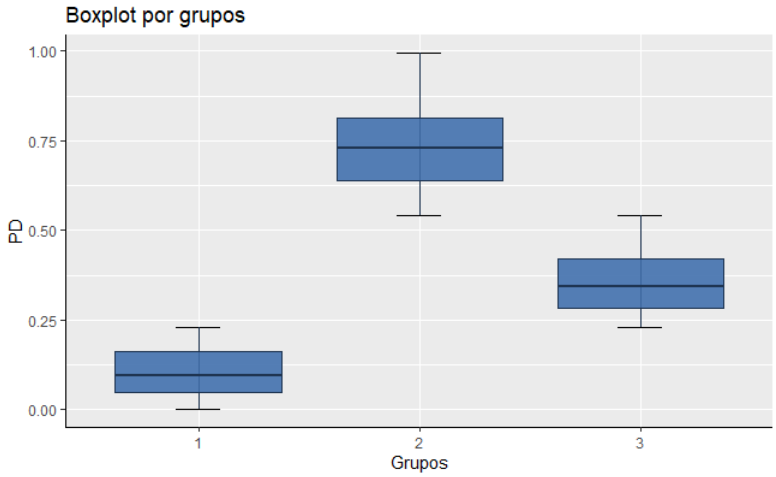


Figura 20: Diagrama de caja de los grupos

Así mismo, se asigna una PD a cada grupo:

GH	PD
1	0.3524894
2	0.9056252
3	0.6166489

15.2.2. Validación de los clústers

Con el fin de evaluar la calidad de agrupamiento de los clusters propuestos en el análisis anterior se emplearan tres estadísticos, Davies-Bouldin, Dunnett T3, test de Bartlett, test Games y Howell.

Índice Davies-Bouldin

DB	0.4911241

El valor del índice es 0,4911241, por lo que se sospecha que nuestros cluster están clasificando correctamente a los datos.

Prueba de Dunnett T3

Después de implementar en test Dunnet T3, se obtuvo un pvalor de 2e-16; es decir, en cada instante detiempo las medias de los grupos son diferentes estad ısticamente, por tanto podemos decir que los 3 grupos sonheterog eneos entre ellos

Bartlett test

Bartlett's	K-squared	df	p-value
GH1 - GH2	**246.25**	**1**	**2.2e-16**
GH1 - GH3	**336.38**	**1**	**2.2e-16**
GH2 - GH3	**969.98**	**1**	**2.2e-16**

como el p_valor es menor a un $\alpha = 0{,}05$, se rechaza la hipótesis nula y, en consecuencia, las varianzas entre el grupos son diferentes.

Test de Games y Howell

group1	group2	estimate	conf.low	conf.high	p.adj	p.adj.signif
GH1	GH2	-0.2547820	-0.2578103	-0.2517538	0.00e+00	****
GH1	GH3	0.3663524	0.3611089	0.3715960	4.33e-08	****
GH2	GH3	-0.6231122	0.6161677	0.6261013	1.11e-08	****

En consecuencia, vemos que los grupos homogéneos son heterogéneos entre sí.

Finalmente calcularemos la tasa de mora de cada grupo y veremos su comportamiento a través del tiempo del modelo PIT.

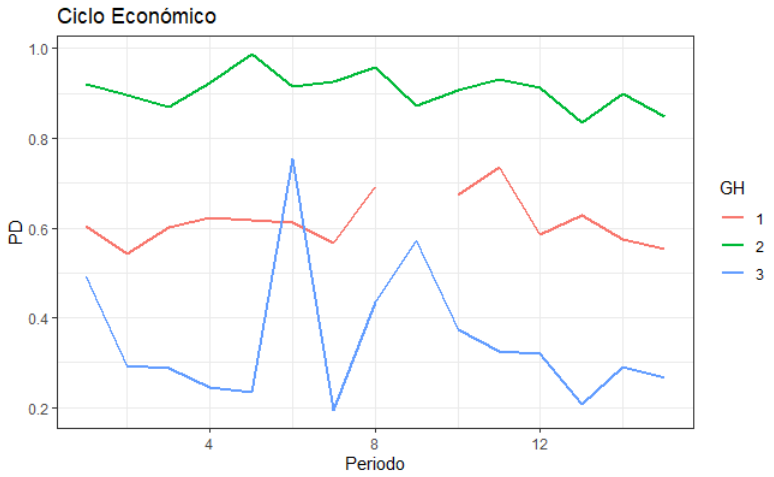


Figura 21: Tasa de mora en el tiempo de los GH

ahora, presentamos un gráfico donde se muestra el porcentaje de cada grupo en cada mes

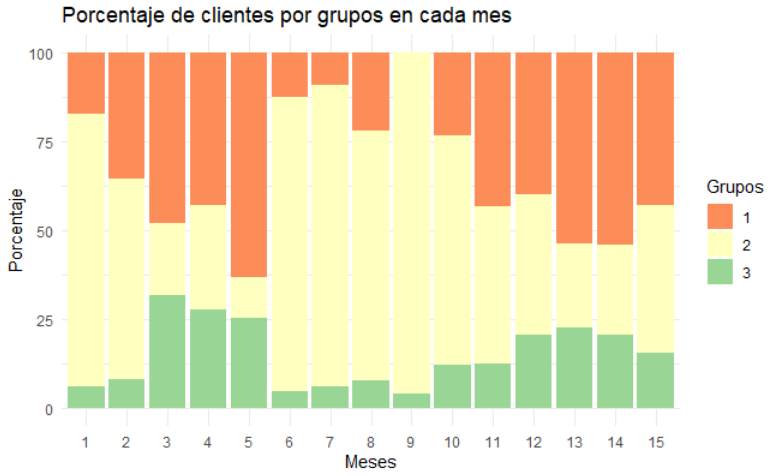


Figura 22: Tasa de mora en el tiempo de los GH

Al ordenar los individuos a lo largo de tiempo considerando el enfoque PIT (con la inclusión de las variables macroeconómicas), se observa en la gráfica precedente una mayor migración de los individuos a comparación de la filosofía TTC. Este resultado era el esperado ya que es más probable que los individuos migren conforme cambian sus características.

16. Alocación de capital

16.1. Pérdida Esperada (PE)

A partir del modelo desarrollado este trabajo tiene el objetivo de calcular la pérdida esperada (PE), esto se logrará con la estimación de los componentes de riesgo que se incorpora en este modelo, incluyen cálculos de la probabilidad de incumplimiento (PD), pérdida en caso de incumplimiento (LGD) y exposición al riesgo de crédito (EAD) estas componentes son estimadas o calculadas de acuerdo a la naturaleza de cada una.
La pérdida esperada es un término estadístico que refleja la probabilidad marginal de que una compañía genere un impago(Funding Circle,2021)[7]; y se expresa de la siguiente manera:

$$PE = PD * EAD * LGD$$

donde:

- **Probabilidad de incumplimiento (PD):**Es la posibilidad de que ocurra el incumplimiento parcial o total de una obligación de pago o el rompimiento de un acuerdo del contrato de crédito, en un período determinado.
- **Exposición dado el incumplimiento (EAD):** Es el valor o saldo al momento de producirse el incumplimiento de los flujos que se espera recibir de las operaciones crediticias, es el valor del saldo insoluto a una fecha de corte determinada.
- **Pérdida en caso de incumplimiento (LGD):**Es una estimación de la parte que realmente se pierde en caso de incumplimiento del cliente.

16.2. Perdida Esperada para la filosofía TTC

16.2.1. LGD

Empezamos con el cálculo del LDG, siguiendo los siguientes pasos:

1. Simulamos una muestra de observaciones de los LGD con una distribución beta de parámetros 0.5
2. Ordenamos los LGD de forma decreciente

$$LGD_C > LGD_B > LGD_A$$

3. Ordenamos los scores por grupo, de forma que el peor grupo tenga los LGD mas altos y el mejor grupo los LGD mas bajos.
4. calculamos la media de los LGD por grupo

A continuación, se muestra la gráfica de la simulación.

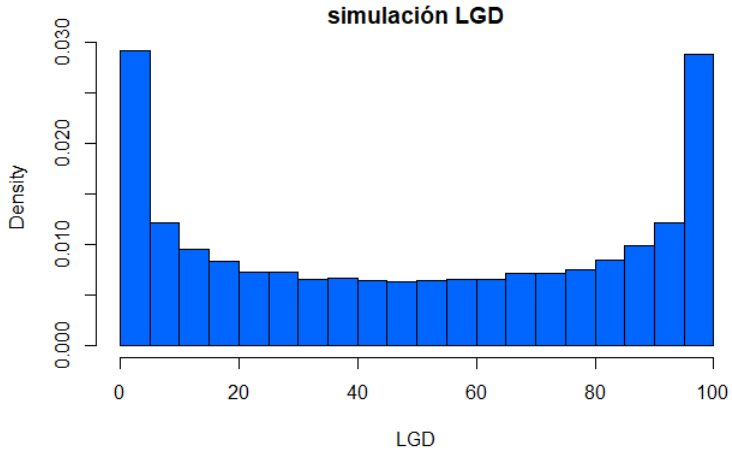


Figura 23: correlación entre las variables

obtenemos los siguientes valores de LDG

GRUPO	A	B	C
LGD	0.1493555	0.5480862	0.8887707

16.2.2. EAD

Por otro lado, el modelamiento para obtener el valor del EAD de cada grupo se han tomado los siguientes valores para el EAD, cumpliendo la condición de grupos en donde,

GRUPO	A	B	C
EAD	100	300	500

$$EAD_A > EAD_B > EAD_C$$

Puesto que si un cliente es considerado un buen pagador formará parte de un mejor grupo y con acceso a un crédito más alto, por ende, el tamaño de la deuda es más alto.

16.2.3. PE para TTC

Una vez realizados lo grupos homogéneos se puede recuperar de ellos las PD para cada grupo, siendo este el valor medio de cada uno, y se han obtenido los siguientes valores:

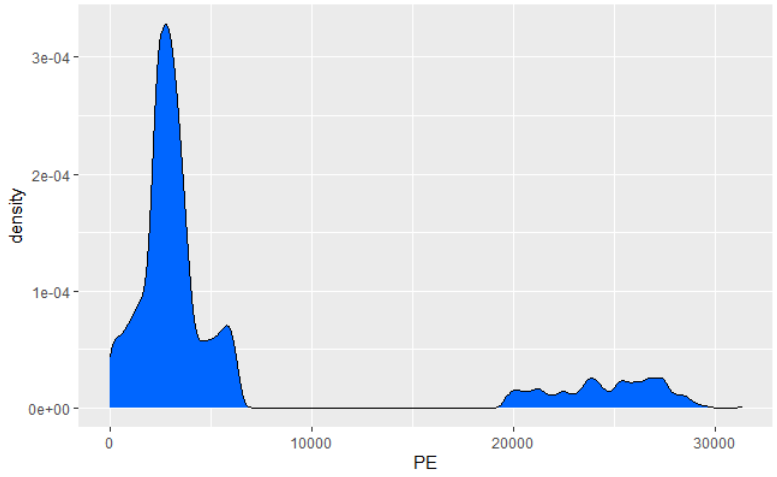


Figura 24: perdida esperada TTC

ahora, calculamos los capitales económicos se presentara en una tabla y la media del Capital Económico

VAR	28151.19
PE de la cartera:	6504.322
Capital económico	21646.87

Finalmente se calcula los capitales económicos de cada tiempo(meses) y se presentara en una tabla y la media del Capital Económico por tiempo

Fecha	Var	PE	Capital
1	23731.17	4296.751	19434.42
2	26650.17	4067.398	22582.78
3	28490.6	9937.357	18553.25
4	28574.51	8838.955	19735.55
5	27933.52	7562.626	20370.89
6	23676.76	3982.721	19694.04
7	22022.73	3210.967	18811.76
8	26535.56	4623.584	21911.98
9	22494.48	3277.146	19217.34
10	27434.37	5625.412	21808.96
11	27973.17	5949.382	22023.78
12	28569.55	7709.556	20860.00
13	28753.33	8599.935	20153.40
14	28300.82	7635.627	20665.20
15	28568.59	6813.367	21755.22

16.3. Perdida Esperada para la filosofía PIT

De la misma forma que el filosofía anterior, realizaremos el cálculo de la simulación de LGD siguiendo los mismos pasos.

obtenemos los siguientes valores de LDG

GRUPO	A	B	C
LGD	0.0230818	0.3535409	0.89261820

16.3.1. EAD

Por otro lado, el modelamiento para obtener el valor del EAD de cada grupo se han tomado los siguientes valores para el EAD, cumpliendo la condición de grupos en donde,

GRUPO	A	B	C
EAD	100	300	500

$$EAD_A > EAD_B > EAD_C$$

Puesto que si un cliente es considerado un buen pagador formará parte de un mejor grupo y con acceso a un crédito más alto, por ende, el tamaño de la deuda es más alto.

16.3.2. PE para TTC

Una vez realizados lo grupos homogéneos se puede recuperar de ellos las PD para cada grupo, siendo este el valor medio de cada uno, y se han obtenido los siguientes valores:

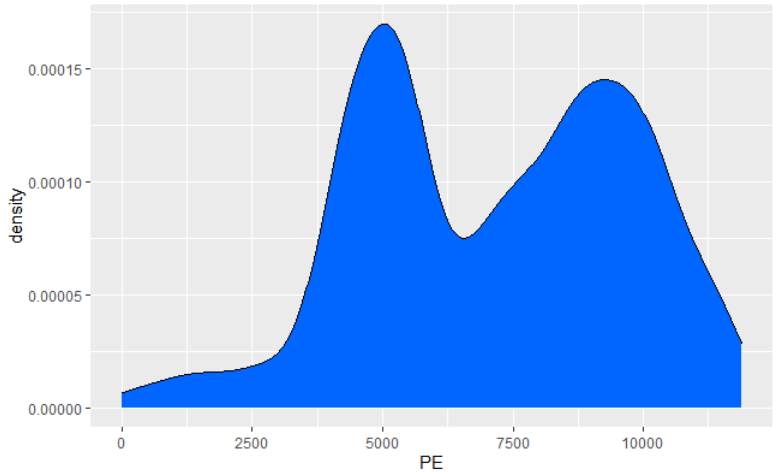


Figura 25: perdida esperada PIT

ahora, calculamos los capitales económicos se presentara en una tabla y la media del Capital Económico

VAR	11628.69
PE de la cartera:	6441.553
Capital económico	5187.138

Finalmente se calcula los capitales económicos de cada tiempo(meses) y se presentara en una tabla y la media del Capital Económico por tiempo

Fecha	Var	PE	Capital
1	11652.58	5415.236	6237.348
2	11507.96	5992.44	5515.519
3	11648.36	7143.364	4504.993
4	11221.35	6857.287	4364.061
5	11676.89	7674.673	4002.217
6	11105.81	5133.082	5972.725
7	11738.59	4319.107	7419.48
8	10634.6	5795.684	4838.914
9	10377.7	4495.181	5882.521
10	11626.43	5710.413	5916.021
11	11588.94	6793.207	4795.731
12	11566.06	6712.447	4853.614
13	11735.78	7482.411	4253.365
14	11684.14	7288.834	4395.303
15	11320.1	6840	4480.095

17. Conclusiones

- El score de crédito permite discrimininar con mayor efectividad a los individuos, respecto de su comportamiento frente a la otorgación de créditos.
- El modelo score es de gran importancia, pues permite identificar los malos y buenos pagadores de créditos, basándose en su información de Marca_Mora, lo cual brinda un beneficio al banco en cuestión de pérdidas relacionadas a cualquier tipo de decisión de concesión de crédito inapropiada.
- Es fundamental realizar un análisis de los requerimientos estadísticos del Comité de Basilea para poder desarrollar de forma adecuada el enfoque de este proyecto, y para entender las prácticas y normativas de supervisión bancaria.
- Se podría obtener una mayor precisión del modelo con una mayor temporalidad de datos registrados disponible.
- Dentro del modelamiento del score de crédito, se puede usar dos filosofías de clasificación PIT y TTC. Ambas se diferencian por el tipo de información de los clientes que usa; por un lado el PIT utiliza tanto variables idiosincráticas como sistemáticas y el TTC únicamente variables idiosincráticas.
- El modelo TTC tiende a ser más volátil y por ende resultó en una curva que seguía el ciclo económico.
- La filosofía TTC no se ve afectado por las fluctuaciones de la economía, ya que, solo usa información idiosincrática de los clientes. Por lo tanto, la PD se mantiene constante, bajo este sistema de clasificación.
- Al clasificar los grupos homogéneos mediante el enfoque TTC se obtuvo un número óptimo de 3 grupos identificados.

Referencias

Referencias

[1] Joaquín Amat. *Regresión logística simple y múltiple*. www.cienciadedatos.net. 2016.

[2] BBVA. *Sistema Bancario*. 2021.

[3] V Berger e Y Zhou. "Kolmogorov-Smirnov Test: Overview". En: *Wiley StatsRef: Statistics Reference Online* (2014), págs. 1-5.

[4] Tim Bock. *¿Qué son los factores de inflación de varianza (VIF)?* https://www.displayr.com/variance-inflation-factors-vifs/.

[5] Julián Cárdenas. *Odd ratio: qué es y cómo se interpreta*. http://networkianos.com/odd-ratio-que-es-como-se-interpreta/. 2015.

[6] D Chorafas. *Managing credit risk, analysing rating and pricing the probability of default*. Londres: Euromoney Institutional Investor PLC, 2000.

[7] Funding Circle. *Pérdida esperada (PE)*. https://www.fundingcircle.com/es/diccionario-financiero/perdida-esperada. 2021.

[8] Georgios Drakos. *Silhouette Analysis vs Elbow Method vs Davies-Bouldin Index: Selecting the optimal number of clusters for KMeans clustering*. GDCoder. https://gdcoder.com/silhouette-analysis-vs-elbow-method-vs-davies-bouldin-index-selecting-the-optimal-number-of-clusters-for-kmeans-clustering/. Mar. de 2020.

[9] David Hamilton. *An introduction to Through the Cycle Public Firm*. Moody´s Analytics. Mayo de 2011.

[10] Jiawei Han, Micheline Kamber y Jian Pei. "DATA MINING Concepts and Techniques". En: Morgan Kaufmann, 2012. Cap. 10 Cluster Analysis: Basic Concepts and Methods.

[11] Lourdes Hernández. *Datos no balanceados. Sobremuestreo, submuestreo y ponderación*. dr. metrics. https://www.doctormetrics.com/datos-no-balanceados/. enero de 2019.

[12] Google IA. "Clasificación: ROC y AUC". En: developers.google. Cap. Clasificación.

[13] Gary King y Zeng Langche. *Logistic Regresssion in Rare Events Data*. Society for Political Methodology. 2001.

[14] S Moreno. *El modelo logit mixto para la construcción de un scoring de crédito*. 2013.

[15] Celia Salcedo. "Estimación de la ocurrencia de incidencias en declaraciones de pólizas de importación". En: Oficina General del Sistema de Bibliotecas y Biblioteca Central UNMSM. Cap. Capítulo 2.

[16] N Siddiqi. *Credit Risk Scorecards*. Wiley, 2006.

[17] J Soler y col. *Gestión de riesgos financieros: Un enfoque práctico para países latinoamericanos*. Grupo Sant. Washington, DC: Banco Interamericano de Desarrollo, 1999.

[18] Carolina Yépez. *Evaluación de riesgo crediticio en una cartera de consumo bajo dos sistemas de clasificación: procíclico y acíclico*. 2019.