
# ESCUELA POLITÉCNICA NACIONAL

## Modelos de Riesgo

---

# Modelo de Credit Scoring

---

*Integrantes:*
Lizeth Moreno
Wagner Salazar
María Belén Rosero

*Profesor:*
Dra. Adriana Uquillas

Quito - Ecuador
12 de enero de 2021

# Índice

# 1. Resumen

Dada las crisis financieras en distintas entidades bancarias, el Comité de Basilea aborda diferentes metodologías de clasificación para evaluar el riesgo de crédito.

En el presente trabajo se construye un modelo de Scoring, a partir, de la metodología Logit considerando dos enfoques PIT y TTC.

Como primer paso se estimó un Modelo Logit para obtener las probabilidades de incumplimiento de un crédito de un cliente de una entidad financiera. Luego para analizar la estabilidad del sistema de clasificación se crearon grupos homogéneos de riesgo.

# 2. Introducción

La concesión de créditos es uno de los principales negocios de las instituciones bancarias, que a su vez puede ocasionar la quiebra de las mismas. Tal es el caso de numerosos bancos europeos que, en la actualidad, están pasando por una delicada situación debido a la creciente tasa de morosidad, obligando así a dichas entidades a incrementar la provisión por insolvencia y eliminado cualquier posibilidad de beneficio e incluso llegando a tener que soportar importantes pérdidas.

Ante esta situación el Comité Basilea emitió recomendaciones y acuerdos en distintos organismos para un control financiero eficiente. Entre estos se introjueron los acuerdos de Basilea y los sistemas de clasificación de riesgo PIT (Point in time) y TTC (Through the cycle).

Este Comité protege los depósitos con el capital de las instituciones financieras y exige que el capital sea calculado para cubrir los riesgos crediticios de la institución. De tal manera, las instituciones financieras se ven en la necesidad de construir modelos de clasificación interna que aseguren con alta probabilidad que el cliente será capaz de hacer frente a sus obligaciones crediticias incorporando así calidad a sus créditos. El análisis equivocado del cliente implicaría un déficit del mismo que provoca carteras morosas, pérdidas e impiden un desarrollo correcto de la institución.

Se propone crear un modelo de scoring de calificación de crédito o credit scoring, denominado así a todo sistema de evaluación crediticia 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 poíticas y macroeconómicas, que van a definir cada observación, es decir, cada solicitud de crédito.

Se desarrolla el modelo de credit scoring con técnicas paramétricas como es el Modelo de Regresión Logística con dos sistemas de clasificación de riesgo PIT(Point in time) y TIC (Through the cycle) para estimar la probabilidad de impago de un cliente.

El sistema de clasificación PIT (Point in time) se centra en la situación de riesgo actual de los deudores, dada la industria predominante y condiciones económicas.

El sistema de clasificación TTC (Through the cycle) enfatiza el conservadurismo y la estabilidad de las estimaciones de riesgo y las transiciones de calificación. Mira más allá de la situación económica inmediata y, en cambio, se centra en el riesgo esperado durante los escenarios de estrés, como un punto muerto en el negocio de la empresa. (The Credit Scoring Toolkit,2007).

## 3. Objetivos

- Determinar el componente de riesgo de la institución financiera.
- Encontrar un modelo de regresión logística para una entidad bancaria, con el cual se pueda tener una respuesta adecuada sobre si se otorga o no un crédito a una persona según la clasificación PIT y TTC
- Crear grupos homogéneos de riesgo.

## 4. Herramientas

Para cumplir con los objetivos propuestos anteriormente y obtener resultados, se lo realizó con el software estadístico R, debido que cuenta con funciones que facilitan los cálculos necesarios. Además de que la base fue proporcionada en Excel, se utilizó este software para obtener gráficos de la creación de grupos.

## 5. Descripción de la base de datos

La base de datos proporcionada es: "bank-additional-full", que consiste en datos de clientes que han recibido un crédito y contiene 21 variables y 41188 observaciones.

A continuación se presentan las variables de dicha base.

| NOMBRE | DESCRIPCIÓN | TIPO |
|---|---|---|
| age | Edad del Cliente | Numérico |
| job | Ocupación del cliente | Categórica |
| marital | Estado Civil | Categórica |
| education | Nivel de educación del cliente | Categórica |
| default | Indica si el cliente tiene un crédito en mora | Categórica |
| housing | Indica si el cliente tiene un crédito hipotecario | Categórica |
| loan | Indica si el cliente tiene préstamos personales | Categórica |
| contact | Tipo de comunicación que se tiene con el cliente | Categórica |
| month | Mes en el que se hizo el último contacto | Categórica |
| day_of_week | Día de la semana en el que se hizo el ultimo contacto | Categórica |
| duration | Duración del último contacto en segundos | Numérico |
| campaign | Número de contactos realizados durante la campaña para este cliente | Numerico |
| pdays | Número de días desde la última vez que se contactó con el cliente. | Numérico |
| previous | Número de contactos realizados antes de la campaña actual | Numérico |
| poutcome | Resultado de la campaña de marketing anterior | Categórica |
| empvarrate | Tasa de variación del empleo- Indicador de cuartiles | Numérico |
| conspriceidx | Índice de precios al consumidor- Indicador mensual | Numérico |
| consconfidx | Índice de confianza del consumidor- Indicador mensual | Numérico |
| euribor3m | Tasa de interés de bancos europeos - Indicador diario | Numérico |
| nremployed | Número de empleados- Indicador cuartiles | Numérico |

Notemos además que de las presentes variables se tomaron en consideración como variables macroeconómicas y variables idiosincráticas las siguientes:

**Variables idiosincráticas:**

- Edad
- Trabajo
- Estado civil
- Educación
- Housing
- Loan
- Contact
- Month
- Day of week
- Duration
- Campaing
- Pdays
- Poutcome
- Nremployed

**Variables macroeconómicas:**

- Empvarrate
- Conspriceidx
- Consconfidx
- Euribor3m

## 5.1. Adición de Variables

Cabe mencionar que para la creación de grupos de riesgo homogéneos se añadieron las variables monthd y prob (probabilidades de incumplimiento), esta última proporcionada por el modelo de regresión logística obtenido.

## 5.2. Definición de la Variable Dependiente

La variable 'y' de incumplimiento indica si el cliente tiene un crédito en mora o no.

$$y = \begin{cases} 1 & \text{si el cliente tiene un crédito en mora} \\ 0 & \text{caso contrario} \end{cases} \tag{1}$$

Definir bien esta variable es un factor importante puesto que es la variable que se va a predecir mediante el modelo de scoring.

Esta variable nos devolverá valores binarios aplicando regresión Logística.

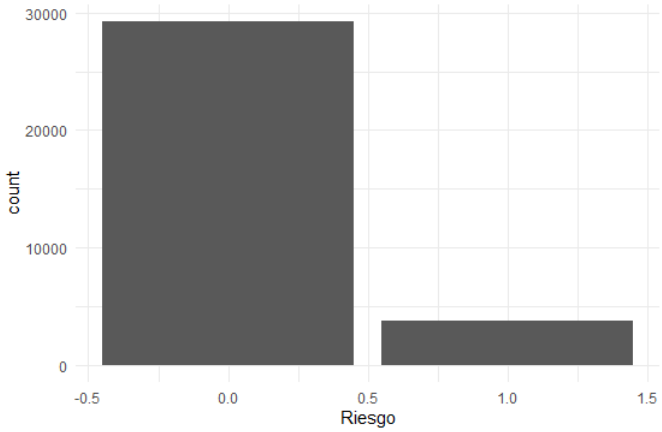



Figura 1: Tasa de clientes que no caen en mora vs caen en mora

En la figura 1, se observa la cantidad de clientes que no tienen un crédito en mora vs la cantidad de clientes que tienen un crédito en mora. Dónde se puede notar que en porcentaje del número de personas que no tienen un crédito en mora es mayor al que si. Además, podemos decir que se describe a un cliente como "buen pagador"si no tiene un crédito en mora.

## 5.3. Tratamiento de la base de datos

### 5.3.1. Partición de la data

La base de datos proporcionada fue dividida de manera aleatoria en un 80 % (entrenamiento) para realizar la modelización y un 20 %(prueba) que corresponde a los datos para la validación del modelo.

### 5.3.2. Selección y Limpieza de datos

En esta sección se depurará la base permitiendonos identificar datos incompletos, incorrectos, inexactos, no pertinentes, etc. De tal manera que podamos substituir, modificar o eliminar estos datos con el fin de obtener una base de datos de calidad y tomar decisiones estratégicas correctas.

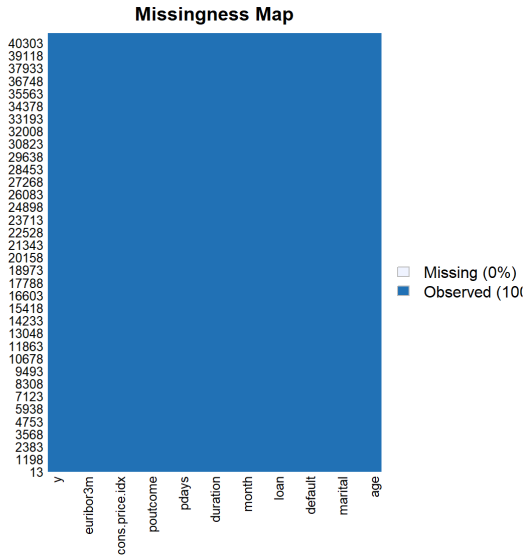



Figura 2: Datos faltantes

Como se observa en la Figura 2, la base de datos proporcionada tiene un 100 % de los datos, es decir, la base esta completa.

Ahora se procede a ver que el tipo de variables este de acuerdo con la información proporcionada:

| VARIABLE | TIPO |
|---|---|
| edad | : num [1:41188] 56 57 37 40 56 45 59 41 24 25 ... |
| trabajo | : chr [1:41188] "housemaidservicesservicesadmin."... |
| marital | : chr [1:41188] "marriedmarriedmarriedmarried"... |
| educación | : chr [1:41188] "basic.4yhigh.schoolhigh.schoolbasic.6y"... |
| predeterminado | : chr [1:41188] "nounknownnono"... |
| vivienda | : chr [1:41188] "nonoyesno"... |
| préstamo | : chr [1:41188] "nononono"... |
| contacto | : chr [1:41188] "telephonetelephonetelephonetelephone"... |
| mes | : chr [1:41188] "maymaymaymay"... |
| día_de_semana | : chr [1:41188] "monmonmonmon"... |
| duración | : num [1:41188] 261 149 226 151 307 198 139 217 380 50 ... |
| campaña | : num [1:41188] 1 1 1 1 1 1 1 1 1 1 ... |
| pdías | : num [1:41188] 999 999 999 999 999 999 999 999 999 999 ... |
| anterior | : num [1:41188] 0 0 0 0 0 0 0 0 0 0 ... |
| poutcome | : chr [1:41188] "nonexistentnonexistentnonexistentnonexistent"... |
| tasa de variación de emp | : chr [1:41188] "1.11.11.11.1"... |
| cons.price.idx | : num [1: 41188] 93994 93994 93994 93994 93994 ... |
| cons.conf.idx: | : chr [1: 41188] 36.4-36.4-36.4-36.4"... |
| euribor3m | : num [1:41188] 4857 4857 4857 4857 4857 ... |
| n. empleado | : num [1:41188] 5191 5191 5191 5191 5191 ... |
| y | : chr [1:41188] "nononono"... |

Se observa que las variables emp.var.rate y cons.conf.idx son numéricas pero estan guardadas como caracter, por lo cual debemos cambiar a tipo numéricas.

# 6. Tratamiento de valores atípicos

En esta sección se identifica las observaciones que numéricamente son distintas al resto de datos, es decir, identificamos valores atípicos porque podrían tener un efecto desproporcionado en los resultados estadísticos.

## 6.1. Variables Continuas

Se analiza la no linealidad y la relación que existe entre la variable continua con la variable respuesta para distinguir que transformación se deberá colocar en las variables para el modelo de regresión .

Cabe mencionar que se utilizarán los siguientes test para los previos análisis:

**Prueba de Lilliefors**
La prueba de Lille.test propuesta en R es una prueba de normalidad basada en la prueba de Kolmogorov-Smirnov dónde se prueban las hipótesis

$$H_0 : \text{los datos provienen de una distribución normal.}$$

$$H_a : \text{los datos no provienen de una distribución normal.}$$

Por otro lado, la prueba de Lilliefors con base en Kolmogorov-Smirnov se aplica para contrastar la hipótesis de normalidad de la población, el estadístico de prueba es la máxima diferencia:

$$D = máx|F_n(x) - F(x)|$$

donde, $F_n(x)$ la función de distribución empírica y $F(x)$ la función de distribución acumulativa de la distribución normal con la media estimada y la varianza estimada.Entonces la prueba, evalúa si la discrepancia máxima es lo suficientemente grande como para ser estadísticamente significativa , lo que requiere el rechazo de la hipótesis nula. Aquí es donde esta prueba se vuelve más complicada que la prueba de Kolmogorov-Smirnov. Debido a que la hipotética F(x) se ha movido más cerca de los datos mediante una estimación basada en esos datos, la discrepancia máxima se ha hecho más pequeña de lo que hubiera sido si la hipótesis nula hubiera destacado solo una distribución normal. Por lo tanto, la "distribución nula"del estadístico de prueba, es decir, su distribución de probabilidad suponiendo que la hipótesis nula es cierta, es estocásticamente más pequeña que la distribución de Kolmogorov-Smirnov. Esta es la distribución de Lilliefors.

El paquete en R que nos facilita la función para este test es "nortest", aplicando lillie.test donde nos refleja el p valor para rechazar o no la hipótesis nula propuesta.

**Prueba de Grubbs**
La prueba de Grubbs (prueba de desviación extrema studentizada) se utiliza para detectar un único valor atípico en un conunto de datos univariado dado que el supuesto de que la variable sigue una distribución normal.

$$H_0 : \text{No hay valores atípicos en el conjunto de datos.}$$

$$H_a : \text{Hay exactamente un valor atípico en el conjunto de datos.}$$

La prueba de Grubbs detecta un valor atípico a la vez. Este valor atípico se elimina del conjunto de datos y la prueba se repite hasta que no se detectan valores atípicos. Su estadístico viene dado por

$$G = \frac{\text{máx}_{i=1,\ldots,N} |Y_i - \bar{Y}|}{s}$$

con $\bar{Y}$ y $s$ que denota la media muestral y la desviación estándar , respectivamente. La función que nos permite realizar este test en R es grubbs.test disponible en el paquete "Outliers"

**Prueba de Tietjen-Moore**

La prueba de Tietjen-Moore es una variación de la prueba de Grubbs pero esta se utiliza cuando se sospecha que puede haber más de un valor atípico. Además para el uso de esta prueba se debe conocer el total de valores atípicos en el conjunto de datos.

A continuación, se sigue con el tratamiento de valores atípicos de las variables de la base:

**EDAD**

En primer lugar, vamos a mostrar mediante un histograma la distribución de la variable continua age, adicionalmente, se presenta la forma de la curva observada y teórica para ver si existe aproximación entre ellas.

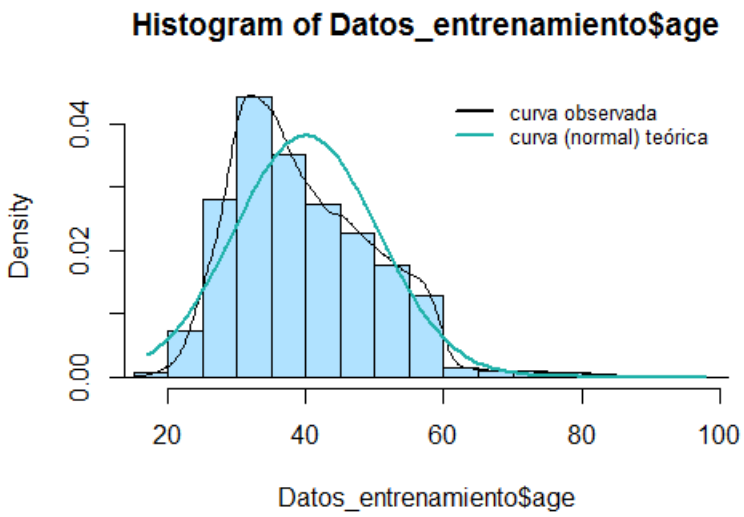



Figura 3: Edad

Evidentemente, a partir de la figura 3 se observa que la distribución de los datos tienden a una distribución normal por lo que se procede a realizar los test ya mencionados para validar la hipótesis de normalidad.

Por tanto, la prueba de lilliefors nos indica que no se rechaza la hipótesis nula, por lo que podemos proceder a aplicar el test de Grubbs para ver la presencia de valores atípicos bajo las hipótesis propuestas, se obtuvo un p valor = 0.00045 lo que implica que se rechaza la hipótesis nula.

Ahora, de la Figura 3 se observó que la relación no es lineal y existe sesgo a partir de la edad de 60 años. En concordancia con lo visto en clases se procede al remplazo de todos los valores a partir del sesgo de 60 por 60 años, y se obtuvo los siguientes resultados:

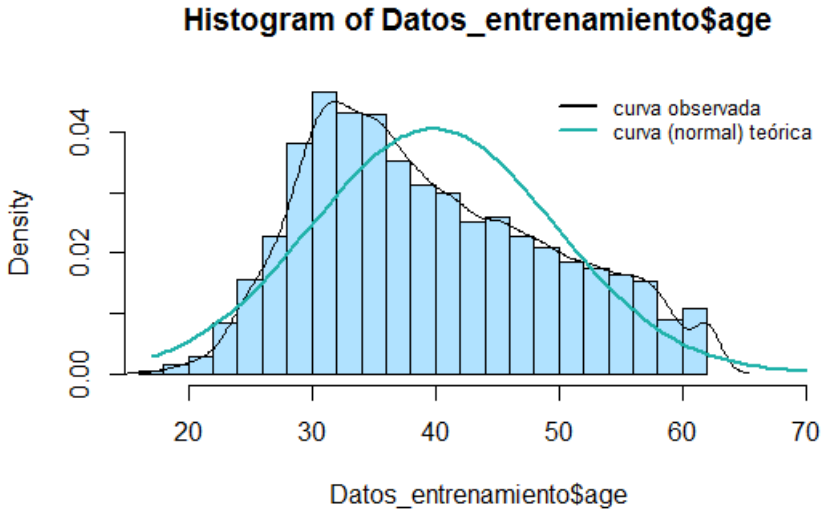
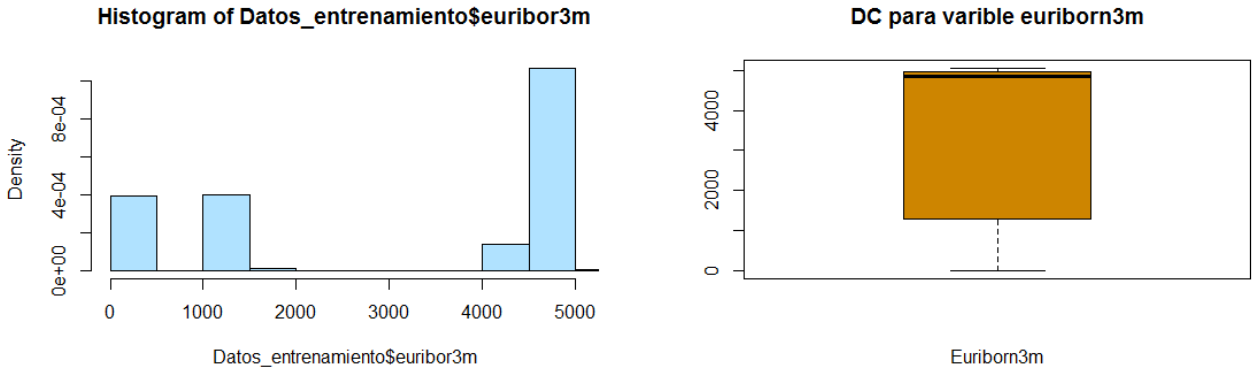



Figura 4: Edad

Se valida la prueba de Lilliefors para verificar la aproximación de los datos a una distribución normal. Se sigue con la aplicación del test de Grubbs, que finalmente nos asegura que ya no hay presencia de valores atípicos, como se puede observar en la figura 4.

## EURIBORN3M

Para esta variable podemos identificar en las siguientes figuras que no existen datos atípicos por lo cuál no es necesario manipular esta variable.



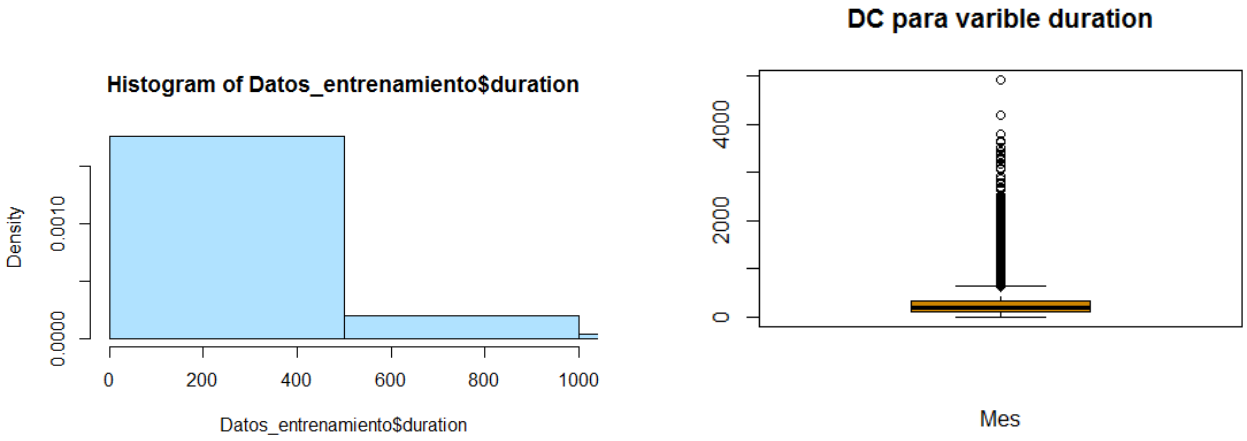

((a)) ((b))

Figura 5: Múltiples imágenes

**Observación:** del mismo modo, para las variables serit y emp.var.rate no se encontró presencia de valores atípicos por lo que no se manipulará dichas variables.

## DURATION

Para esta variable se obtiene el histograma y el diagrama de caja en donde se observa (Figura 9) que no se tiene una distribución conocida y la presencia de datos atípicos:




((a)) Duration ((b)) Duration

Figura 6: Histograma y Diagrama de Caja de la variable duration.

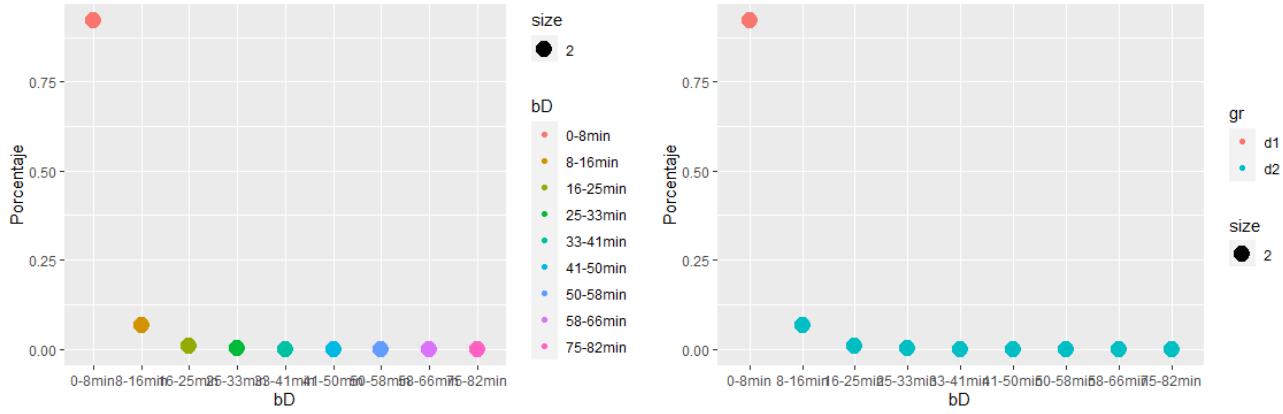
Verificamos la relación existente entre la variable 'duration' y el número de ceros sobre el total,y se procede a establecer dos categorías basada en la relación propuesta, es decir las llamadas menores a 8 min representan un 92 % del número de ceros sobre el total, por lo cual para corregir los valores atípicos de esta variable se procedió a realizar 2 cluster, uno para las llamadas menores a 8 min y otro para las mayores a 8 min.

$$\text{d1} -> \text{[0-8min]}$$
$$\text{d2} -> \text{[9-82min]}$$

La siguiente tabla muestra la suma, porcentajes y los respectivos intervalos de tiempo a los que se asignará los grupos.

| **Suma** | **Porcentaje** | **bD** | **gr** |
|---|---|---|---|
| 26968 | 0.92227 | 0-8min d1 | d1 |
| 1958 | 0.06696 | 8-16min | d2 |
| 246 | 0.00841 | 16-25min | d2 |
| 46 | 0.00157 | 25-33min | d2 |
| 9 | 0.00031 | 33-41min | d2 |
| 2 | 0.00007 | 41-50min | d2 |
| 6 | 0.00021 | 50-58min | d2 |
| 2 | 0.00007 | 58-66min | d2 |
| 1 | 0.00003 | 75-82min | d2 |

Con la información presentada en la tabla se presenta la siguiente figura 7 donde se observa la creación de los grupos.




((a)) Antes del cluster ((b)) Después del cluster

Figura 7: Múltiples imágenes

Una vez que se ha tratado los valores atípicos de la variable 'duration', se identifican los grupos para la variable 'y', como se muestra a continuación:

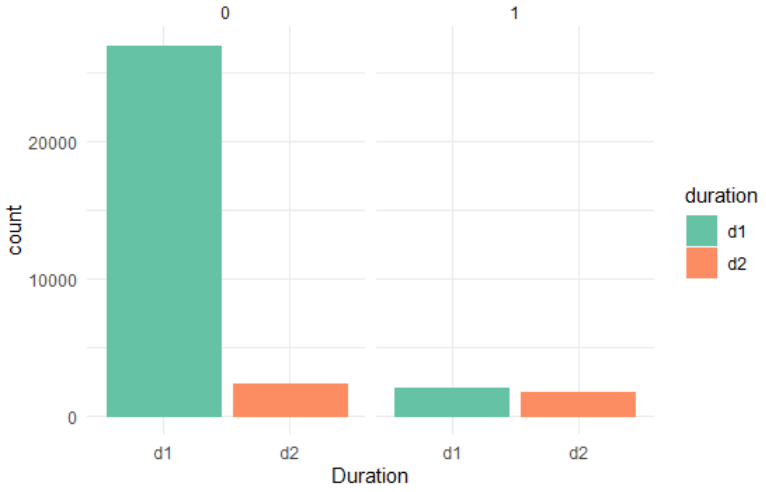
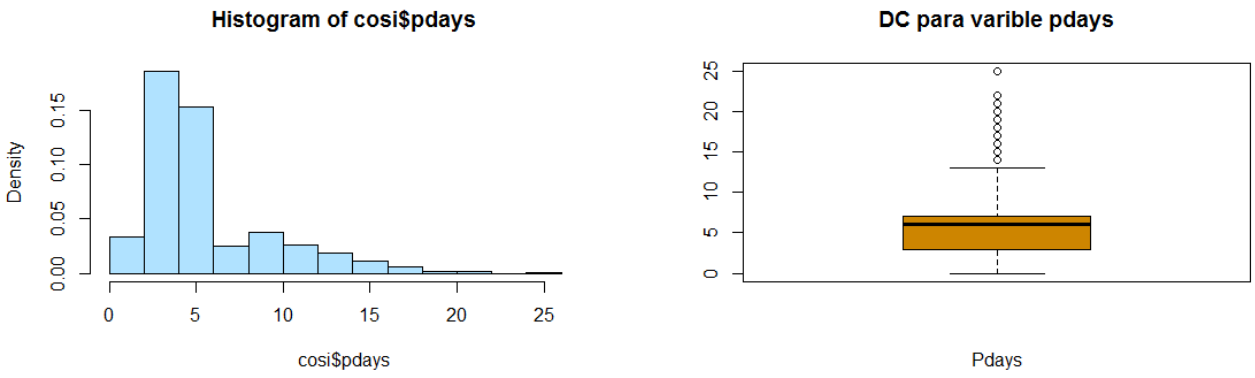



Figura 8: Duration

**Observación:** el proceso que se presentó para generar grupos en la variable Duration, se implementó en las siguientes variables: monthd (3 grupos), campaing (2 grupos), previous (2 grupos), const.price.idx (2 grupos) y cons.conf.idx (2 grupos).

## PDAYS

El análisis de esta variable será diferente, en primer lugar quitaremos los valores 999 que representa que un cliente no ha sido contactado, posterior a eso con los datos restantes realizaremos el análisis de atípicos y de ser el caso la realización de clusters como en las variables anteriores. Una vez retirados los valores 999 tenemos las siguientes figuras:




((a)) ((b))

Figura 9: Múltiples imágenes

Gracias al histograma y el diagrama de caja, dado que no se tiene una distribuión conocida y se tiene datos atípicos; acorde a los visto en clase podemos verficar la relación existente entre la variable pdays y el número de ceros sobre el total, es decir, el número de días desde el último contacto menor a 14 representan un 94 % del número de ceros sobre el total, por lo cual para corregir los atípicos de esta variable se procedió a realizar 3 cluster, uno para el número de días menores a 14, otro para los mayores a 14 hasta 27 días y otros para quienes no han sido contactados.

$$\begin{aligned} \text{d1} &-> [0\text{-}14] \\ \text{d2} &-> [14\text{-}27] \\ \text{d3} &-> \text{no contactado} \end{aligned}$$

Con lo cual la variable pdays en relación a la variable 'y' se observa a continuación:

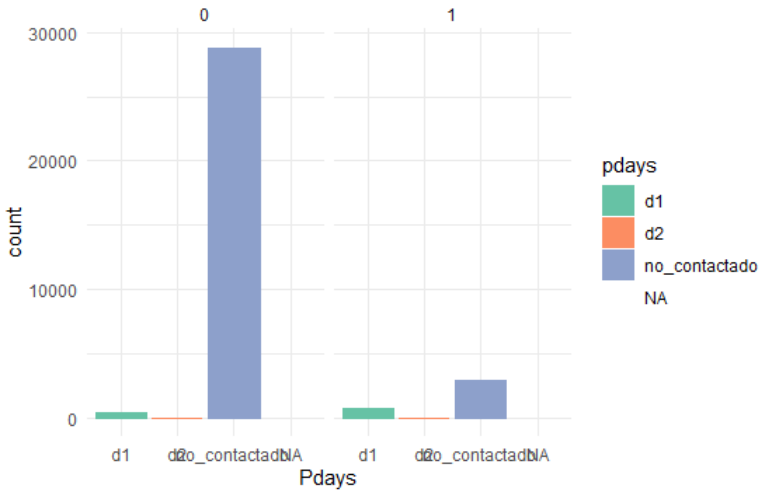



Figura 10

## 6.2. Variables Categóricas

Para esta sección, las variables categóricas 'job', 'education' y 'month' se recategorizaron con el fin de agruparlas y reducir el número de categorías o cardinalidad de ellas, tomando en cuenta la variable

dependiente “y”.

**JOB**

Como se puede observar en la figura 11, se puede notar la presencia de muchas categorías para esta variable.

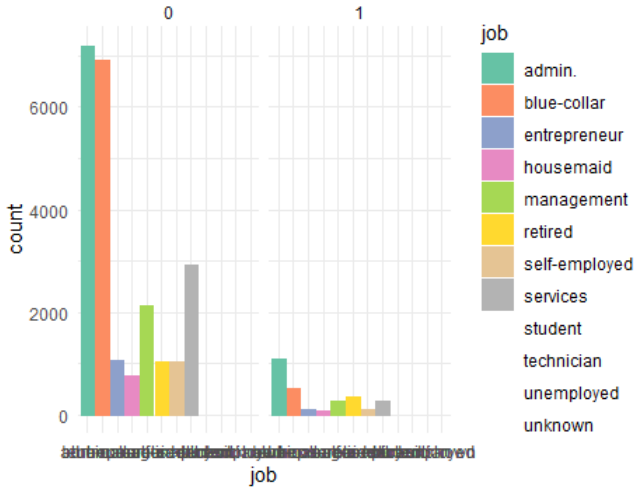
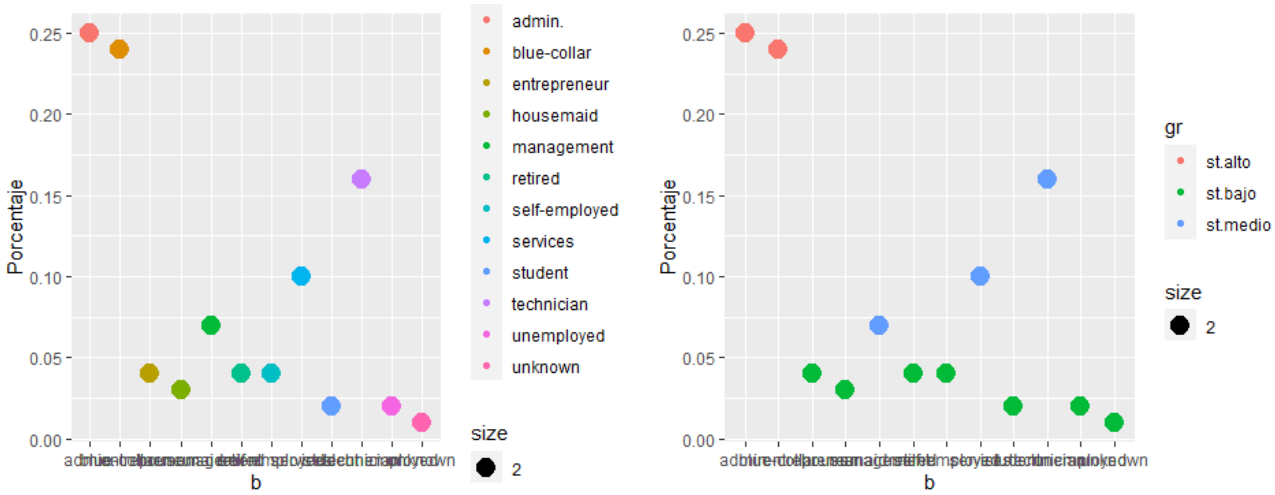



Figura 11

Para reducir la cardinalidad se crearán clusters verificando la cantidad de ceros sobre el total como en la sección anterior.

Crearemos clusters según la siguiente estructura:

$$\text{status alto} -> [0.25\text{-}0.17]$$
$$\text{status medio} -> [0.16\text{-}0.08]$$
$$\text{status bajo} -> [0.07\text{-}0.00]$$




((a)) ((b))

Figura 12: Múltiples imágenes

Con lo cual se ha reducido la cardinalidad a 3, obteniendo la figura 13.


Figura 13

**EDUCATION**

Con el mismo proceso, se recategorizó esta variable y se redujó la cardinalidad de la misma a 3. Figura 14.

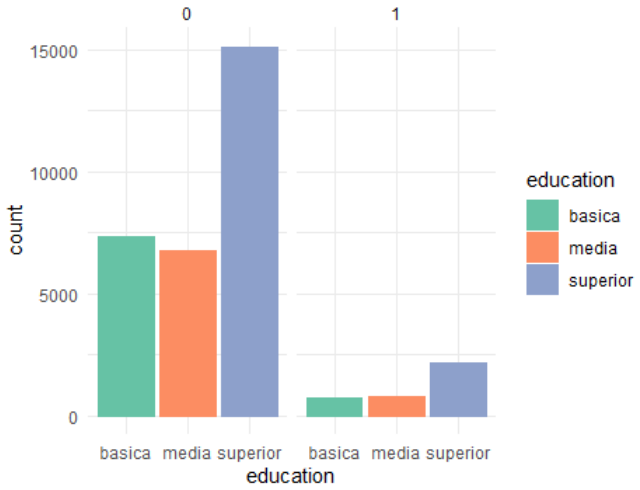



Figura 14

**MONTH**

Se trató esta variable bajo el mismo método obteniendo una nueva categorización como se presenta en la figura 15.

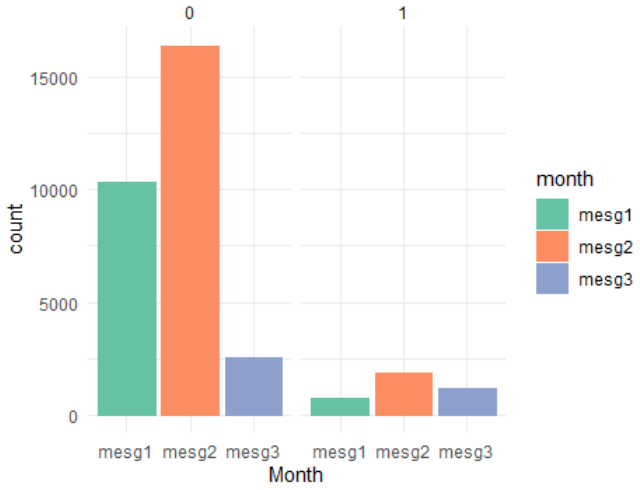



Figura 15

**Observación:** las variables 'marital', 'housing', 'loan', 'contact' y 'day_of_week' no fueron recategorizadas, puesto que su cardinalidad es baja.

# 7. Creación del Modelo Scoring

En esta sección se propone la creación de un modelo de credit scoring, donde el objetivo de este modelo es obtener una probabilidad de mora para que un cliente sea asignado como un buen pagador o un mal pagador. Existen varias técnicas para la modelización del riesgo de crédito, pero se utilizará un modelo de regresión Logística.

## 7.1. Modelo LOGIT

El modelo logit es un modelo de regresión típica $Y = f(X + \epsilon)$, en el que la variable respuesta 'y' es dicotómica, es decir, toma los valores de 0 y 1, y las variables predictivas en este caso son continuas y categóricas.
El modelo Logit, se define a partir de la siguiente función de distribución:

$$P(Y = 1/X_i) = \frac{1}{1 + e^{-Z_i}}$$

Donde:

$$Z_i = \beta_0 + \beta_1 X_i + \epsilon$$

y las variables se definen así:

- $Yi = 1 -> $ Bueno
- $Yi = 0 -> $ Malo
- $Xi -> $ Ingreso de Cliente
- $P(Yi = 1/Xi) -> $ Probabilidad de ser Bueno, explicado por la variable Xi
- $Z_i -> $ Exponente del exponencial que es una regresión líneal

- $\beta_o -> $ Intercepto de la Curva (Parámetro a estimar)
- $\beta_1 -> $ Pendiente de la Curva (Parámetro a estimar)
- $\epsilon -> $ Error
- $i = 1, 2, 3, ...N -> $ Indice de diferenciación de variables

## 7.2. Modelo PIT

### 7.2.1. Selección de las Variables

Para la creación de nuestro modelo se selecciona las variables de acuerdo a las pruebas estadísticas mencionadas en clases como son: Information Value, Estadístico KS y Multicolinealidad.

### 7.2.2. Information Value (IV)

En nuestro proceso de análisis previo de los datos se deben elegir las variales predictorias apropiadas. Así, Information Value es una de las técnicas más útiles para seleccionar variables de tipo categórico importantes en un modelo predictivo. Nos ayuda a clasificar las variables en función de su valor predictivo.

El IV se calcula utilizando la siguiente fórmula:

$$IV = \sum((\text{Distribución buena} - \text{Distribución Mala}) * Ln(\frac{\text{Distribución buena}}{\text{Distribución mala}}))$$

Donde;

- Distribución buena se refiere al porcentaje de valores que da como resultado el "valor a predecir"deseado para la variable dependiente
- Distribución mala es el porcentaje de valores dentro de cada grupo que no es el "valor para predecir"

La siguiente tabla proporciona una regla estándar para usar Information Value para comprender el poder predictivo de cada variable.

| **Poder predictivo** | **Information Value (IV)** |
|---|---|
| Sospechoso | <0.5 |
| Fuerte | 0.3 - 0.5 |
| Medio | 0.1 - 0.3 |
| Débil | 0.02 - 0.1 |
| Inútil | <0.02 |

Las variables que se tomarán en cuenta serán aquellas con un valor predictivo medio y fuerte.

Por último, para calcular los valores predictivos infor_value se utilizó el paquete 'Information' y la función 'iv' aplicado a las variables categóricas de nuestra base de datos .

### 7.2.3. Estadística KS

La prueba Estadística KS(Kolmogórov-Smirnov) es un tipo de prueba no paramétria que se usa para evaluar el error o 'bondad de ajuste' en el ajuste de curvas.

El estadístico de Kolmogorov- Smirnov para una función de distribución acumulativa dada F(x) es:

$$D_n = \sup_x |F_n(x) - F(x)|$$

Este estadístico se utilizó para seleccionar de manera correcta las variables de nuestro modelo. El modelo se dirá eficaz si el estadístico KS es grande, debido a que la variable analizada tendrá mayor poder predictivo.

En este documento, haciendo uso del software R se calcula el estadístico KS utilizando el paquete con la función aplicado a las variables continuas.

Los resultados usando las metodologías vistas anteriormente para todas las variables de nuestra base se resumen en la siguiente tabla:

| | **Variable** | **Test** |
|---|---|---|
| 1 | euribor3m | 1.518 |
| 2 | serit | 1.291 |
| 3 | nr.employed | 1.214 |
| 4 | monthd | 1.081 |
| 5 | emp.var.rate | 1.081 |
| 6 | duration | 0.862 |
| 7 | pdays | 0.559 |
| 8 | poutcome | 0.559 |
| 9 | cons.conf.idx | 0.486 |
| 10 | month | 0.404 |
| 11 | previous | 0.291 |
| 12 | age | 0.258 |
| 13 | contact | 0.254 |
| 14 | job | 0.052 |
| 15 | cons.price.idx | 0.048 |
| 16 | campaign | 0.038 |
| 17 | marital | 0.030 |
| 18 | education | 0.024 |
| 19 | day_of_week | 0.008 |
| 20 | housing | 0.001 |
| 21 | loan | 0.001 |

Los resultados nos indican que las variables que se toman en consideración son: euriborn3m, serit, nr.employed, monthd, emp.var.rate, duration, pdays, poutcome, cons.conf.idx, month, previous, age y contact porque presentan un valor predictivo entre 0.01-0.3 lo que implica que poseen un poder predictivo medio y fuerte. Las otras variables serán eliminadas por presentar poder predictivo inútil y

sospechoso además de las variables seleccionadas bajo el criterio KS, respectivamente para las variables de la tabla presentada.

### 7.2.4. Modelo LOGIT PIT

#### MODELO INICIAL

Ahora, aplicando en el software estadístico R la función glm del paquete MASS para ajustar modelos lineales generalizados se tiene el siguiente resultado de nuestro modelo inicial:

| | **Estimate** | **Std.Error** | **z value** | **Pr(>—z—)** |
|---|---|---|---|---|
| **(Intercept)** | 1.060e+01 | 4.026e+00 | 2.632 | 0.00848 ** |
| age | -1.439e-03 | 2.410e-03 | -0.597 | 0.55050 |
| serit | 8.059e-01 | 3.468e-01 | 2.324 | 0.02013 * |
| euribor3m | 2.011e-05 | 2.234e-05 | 0.900 | 0.36805 |
| nr.employed | -1.638e-03 | 7.933e-04 | -2.065 | 0.03890 * |
| emp.var.rate | -1.624e-01 | 3.620e-02 | -4.485 | 7.28e-06 *** |
| job_st.alto | -3.262e-03 | 5.195e-02 | -0.063 | 0.94993 |
| job_st.bajo | 1.100e-01 | 6.148e-02 | 1.789 | 0.07364 . |
| job_st.medio | NA | NA | NA | NA |
| education_basica | -6.892e-02 | 5.719e-02 | -1.205 | 0.22815 |
| education_media | -3.671e-02 | 5.671e-02 | -0.647 | 0.51737 |
| education_superior | NA | NA | NA | NA |
| marital_divorced | -5.450e-02 | 4.314e-01 | -0.126 | 0.89947 |
| marital_married | -6.450e-02 | 4.268e-01 | -0.151 | 0.87986 |
| marital_single | 4.093e-02 | 4.278e-01 | 0.096 | 0.92377 |
| marital_unknown | NA | NA | NA | NA |
| contact_cellular | 4.375e-01 | 7.346e-02 | 5.956 | 2.59e-09 *** |
| contact_telephone | NA | NA | NA | NA |
| monthd_gr1 | -3.141e+00 | 3.609e-01 | -8.703 | <2e-16 *** |
| monthd_gr2 | -2.653e+00 | 3.057e-01 | -8.678 | <2e-16 *** |
| monthd_gr3 | NA | NA | NA | NA |
| month_mesg1 | -2.588e-01 | 1.272e-01 | -2.035 | 0.04189 * |
| month_mesg2 | -3.921e-01 | 7.747e-02 | -5.062 | 4.15e-07 *** |
| month_mesg3 | NA | NA | NA | NA |
| duration_d1 | -3.182e+00 | 5.429e-02 | -58.610 | <2e-16 *** |
| duration_d2 | NA | NA | NA | NA |
| campaign_d1 | 2.287e-01 | 9.959e-02 | 2.296 | 0.02166 * |
| campaign_d2 | NA | NA | NA | NA |
| previous_gr1 | -4.205e-01 | 2.412e-01 | -1.744 | 0.08121 . |
| previous_gr2 | NA | NA | NA | NA |
| poutcome_failure | -9.066e-01 | 2.308e-01 | -3.928 | 8.55e-05 *** |
| poutcome_nonexistent | NA | NA | NA | NA |
| poutcome_success | NA | NA | NA | NA |
| pdays_d1 | 1.091e+00 | 2.315e-01 | 4.713 | 2.44e-06 *** |
| pdays_d2 | 4.277e-01 | 3.459e-01 | 1.236 | 0.21632 |
| pdays_no_contactado | NA | NA | NA | NA |
| pdays_NA | NA | NA | NA | NA |
| cons.price.idx_d1 | -6.212e-01 | 1.125e-01 | -5.520 | 3.40e-08 *** |
| cons.price.idx_d2 | NA | NA | NA | NA |
| cons.conf.idx_alto | -7.794e-02 | 1.085e-01 | -0.718 | 0.47268 |
| cons.conf.idx_bajo | -4.790e-01 | 1.073e-01 | -4.464 | 8.06e-06 *** |
| cons.conf.idx_media | NA | NA | NA | NA |

Como podemos notar en la tabla anterior, el modelo inicial presenta NAs.

El segundo modelo se forma con la eliminación de las variables anteriores que presentaban NAs pero en el siguiente orden:

- euriborn3m
- job_st.alto .
- marital_married .
- marital_divorced .
- pdays_d2 .
- age .
- nr.employed .
- cons.conf.idx_alto .
- educacion basica .
- education_media .
- job_st.bajo .
- month_mesg1 .

**MODELO 2**

| | **Estimate** | **Std. Error** | **z value** | **Pr(>—z—)** |
|---|---|---|---|---|
| **(Intercept)** | 2.28370 | 0.25730 | 8.876 | <2e-16 *** |
| serit | 1.34469 | 0.25038 | 5.370 | 7.85e-08 *** |
| emp.var.rate | -0.16526 | 0.02346 | -7.043 | 1.88e-12 *** |
| marital_single | 0.12717 | 0.04669 | 2.724 | 0.00645 ** |
| contact_cellular | 0.42429 | 0.07274 | 5.833 | 5.45e-09 *** |
| monthd_gr1 | -3.88675 | 0.22143 | -17.553 | <2e-16 *** |
| monthd_gr2 | -3.22981 | 0.21697 | -14.886 | <2e-16 *** |
| month_mesg2 | -0.44347 | 0.05774 | -7.680 | 1.59e-14 *** |
| duration_d1 | -3.18143 | 0.05414 | -58.765 | <2e-16 *** |
| campaign_d1 | 0.22931 | 0.09944 | 2.306 | 0.02111 * |
| previous_gr1 | -0.53846 | 0.22011 | -2.446 | 0.01443 * |
| poutcome_failure | -0.99932 | 0.21365 | -4.677 | 2.90e-06 *** |
| pdays_d1 | 0.99907 | 0.21208 | 4.711 | 2.47e-06 *** |
| cons.price.idx_d1 | -0.66312 | 0.10687 | -6.205 | 5.47e-10 *** |
| cons.conf.idx_bajo | -0.41444 | 0.07545 | -5.493 | 3.95e-08 *** |

**Validación** Para la validación del modelo 2 se plantean las siguiente hipótesis

$H_0$: Modelo no significativo

$H_a$: Modelo significativo

El estadístico de deviance nos arroja un $p - valor = 0$

Dado que el p-valor obtenido es menor a 0.05 se rechaza $H_0$ por lo que el modelo 2 es significativo.

### 7.2.5. Multicolinealidad

Multicolinealidad es una relación de dependencia lineal fuerte entre la variable 'y' y dos o más variables explicativas.

Para hallar multicolinealidad se analiza la matriz de correlaciones de todas las variables predictorias.

Se aplicará Multicolinealidad para medir el desempeño del modelo y que no exista alta correlación entre las variables explicativas. Los resultados se muestran en la siguiente matriz de correlaciones:

| | **serit** | **emp.var.rate** | **marital_single** | **contact_cellular** | **monthd_gr1** |
|---|---|---|---|---|---|
| **serit** | 1.00000000 | 0.374662969 | -0.027262901 | 0.26870910 | 0.253421780 |
| **emp.var.rate** | 0.37466297 | 1.000000000 | -0.101734945 | -0.39077969 | 0.180604582 |
| **marital_single** | -0.02726290 | -0.101734945 | 1.000000000 | 0.07171394 | 0.003485707 |
| **contact_cellular** | 0,26870910 | -0,390779694 | 0.071713941 | 1.00000000 | -0.178473640 |
| **monthd_gr1** | 0.25342178 | 0.180604582 | 0.003485707 | -0.17847364 | 1.000000000 |
| **monthd_gr2** | 0.31384494 | 0.231112678 | -0.063463243 | 0,06565036 | -0,768622059 |
| **month_mesg2** | 0.23020992 | 0.404842876 | -0.009322899 | 0.20784401 | -0.438540613 |
| **duration_d1** | -0.01390043 | 0.004165956 | -0.004573155 | -0.02479998 | -0.017879241 |
| **campaign_d1** | -0.07807117 | -0.110033248 | 0,005126185 | 0.05755333 | -0.036168301 |
| **previous_gr1** | 0.20860179 | 0.473468903 | -0.044628562 | -0.24128255 | 0.126604914 |
| **poutcome_failure** | -0,05487967 | -0,381964191 | 0.027180723 | 0.20644679 | -0.058222351 |
| **pdays_d1** | NA | NA | NA | NA | NA |
| **cons.price.idx_d1** | 0,16039594 | -0,035802319 | -0.025218339 | 0,16758664 | -0,303867127 |
| **cons.conf.idx_bajo** | -0.58049955 | -0.500236113 | 0.032130811 | 0,13323386 | -0,257187172 |

| | **monthd_gr2** | **mes_mesg2** | **duration_d1** | **campaign_d1** | **previous_gr1** |
|---|---|---|---|---|---|
| **serit** | 0.31384494 | 0,230209919 | -0,013900432 | -0,078071167 | 0,208601795 |
| **emp.var.rate** | 0.23111268 | 0.404842876 | 0,004165956 | -0,110033248 | 0,473468903 |
| **marital_single** | -0.06346324 | -0.009322899 | -0.004573155 | 0,005126185 | -0,044628562 |
| **contact_cellular** | 0.06565036 | 0.207844014 | -0.024799975 | 0,057553328 | -0,241282549 |
| **monthd_gr1** | -0,76862206 | -0,438540613 | -0,017879241 | -0,036168301 | 0,126604914 |
| **monthd_gr2** | 1.00000000 | 0.452582650 | 0.016645673 | -0.015329088 | 0.107312384 |
| **month_mesg2** | 0.45258265 | 1.000000000 | 0,007422685 | -0,069792877 | 0.155971802 |
| **duration_d1** | 0.01664567 | 0.007422685 | 1.000000000 | -0.020424944 | -0.005826817 |
| **campaign_d1** | -0.01532909 | -0.069792877 | -0.020424944 | 1.000000000 | -0.062031069 |
| **previous_gr1** | 0.10731238 | 0,155971802 | -0,005826817 | -0,062031069 | 1.000000000 |
| **poutcome_failure** | -0,05173851 | -0,154644309 | 0.018532905 | 0,047149949 | -0,850746632 |
| **pdays_d1** | NA | NA | NA | NA | NA |
| **cons.price.idx_d1** | 0.39534011 | 0.277333898 | 0.002421372 | 0.055612448 | -0.056256036 |
| **cons.conf.idx_bajo** | -0.20360988 | -0.028757196 | 0,002728049 | 0,053997271 | -0,235594070 |

| | poutcome_failure | pdays_d1 | cons.price.idx_d1 | cons.conf.idx_bajo |
|---|---|---|---|---|
| **serit** | -0,05487967 | NA | 0.160395943 | -0.580499554 |
| **emp.var.rate** | -0,38196419 | NA | -0.035802319 | -0.500236113 |
| **marital_single** | 0.02718072 | NA | -0.025218339 | 0.032130811 |
| **contact_cellular** | 0.20644679 | NA | 0.167586637 | 0.133233857 |
| **monthd_gr1** | -0.05822235 | NA | -0.303867127 | -0.257187172 |
| **monthd_gr2** | -0,05173851 | NA | 0.395340107 | -0.203609882 |
| **month_mesg2** | -0,15464431 | NA | 0.277333898 | -0.028757196 |
| **duration_d1** | 0.01853290 | NA | 0.002421372 | 0.002728049 |
| **campaign_d1** | 0.04714995 | NA | 0.055612448 | 0.053997271 |
| **previous_gr1** | -0,85074663 | NA | -0.056256036 | -0.235594070 |
| **poutcome_failure** | 1.00000000 | NA | 0.089068402 | 0.124651071 |
| **pdays_d1** | N / A | 1 | NA | NA |
| **cons.price.idx_d1** | 0.08906840 | NA | 1.000000000 | -0.080495153 |
| **cons.conf.idx_bajo** | 0.12465107 | NA | -0.080495153 | 1.000000000 |

La tabla anterior nos indica que existe una alta correlación entre alguas variables, lo que nos da un indicio de multicolinealidad.

Para mejor entendimiento de la correlación de las variables se lo puede visualizar en la figura 16.

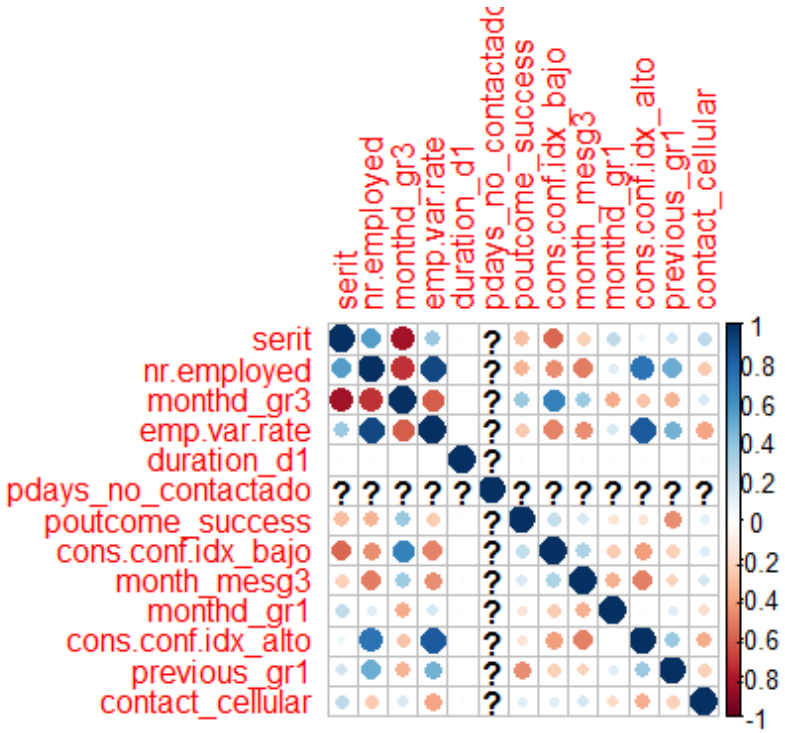



Figura 16: Gráfico de correlación de las variables

### 7.2.6. Factor de varianza inflada (VIF)

Revisamos entonces el factor de influencia de varianza utlizando la función 'vif' del sotfware R para cada variable del modelo PIT y se obtiene el siguiente resultado:

| serit | emp.var.rate | marital_single |
|---|---|---|
| 20.036580 | 3.513371 | 1.020391 |
| **contact_cellular** | **monthd_gr1** | **monthd_gr2** |
| 1.864718 | 22.590092 | 20.629302 |
| **month_mesg2** | **duration_d1** | **campaign_d1** |
| 1.733553 | 1.418356 | 1.023642 |
| **previous_gr1** | **poutcome_failure** | **pdays_d1** |
| 17.021461 | 10.678932 | 7.969014 |
| **cons.price.idx_d1** | **cons.conf.idx_bajo** | |
| 1.256928 | 1.797619 | |

Según los valores obtenidos estamos frente a la existencia de multicolinealidad pues hay 5 variables que tienen un valor superior a 10, entonces procedemos a eliminar la que tiene mayor valor (month_gr1).

**MODELO 3**

| | Estimate Std. | Error | z valor | Pr (>— z —) |
|---|---|---|---|---|
| **(Intercept)** | 1.58427 | 0.24450 | 6.480 | 9.20e-11 *** |
| **serit** | -2.74680 | 0.09208 | -29.829 | <2e-16 *** |
| **emp.var.rate** | -0.34547 | 0.01965 | -17.581 | <2e-16 *** |
| **marital_single** | 0.14801 | 0.04637 | 3.192 | 0.00141 ** |
| **contact_cellular** | 1.10757 | 0.06333 | 17.487 | <2e-16 *** |
| **monthd_gr2** | 0.52290 | 0.05921 | 8.831 | <2e-16 *** |
| **duration_d1** | -3.08132 | 0.05267 | -58.504 | <2e-16 *** |
| **campaign_d1** | 0.10966 | 0.09758 | 1.124 | 0.26110 |
| **previous_gr1** | -0.88237 | 0.21453 | -4.113 | 3.91e-05 *** |
| **poutcome_failure** | -1.28871 | 0.20852 | -6.180 | 6.40e-10 *** |
| **pdays_d1** | 0.81645 | 0.20713 | 3.942 | 8.09e-05 *** |
| **cons.price.idx_d1** | -0.89150 | 0.10308 | -8.648 | <2e-16 *** |
| **cons.conf.idx_bajo** | -0.51242 | 0.07566 | -6.773 | 1.26e-11 *** |

Lo que hace que month_mesg2 deje de ser significativo.

| | Estimate Std. | Error | z value | Pr(>—z—) |
|---|---|---|---|---|
| **(Intercept)** | 1.58427 | 0.24450 | 6.480 | 9.20e-11 *** |
| **serit** | -2.74680 | 0.09208 | -29.829 | <2e-16 *** |
| **emp.var.rate** | -0.34547 | 0.01965 | -17.581 | <2e-16 *** |
| **marital_single** | 0.14801 | 0.04637 | 3.192 | 0.00141 ** |
| **contact_cellular** | 1.10757 | 0.06333 | 17.487 | <2e-16 *** |
| **monthd_gr2** | 0.52290 | 0.05921 | 8.831 | <2e-16 *** |
| **duration_d1** | -3.08132 | 0.05267 | -58.504 | <2e-16 *** |
| **campaign_d1** | 0.10966 | 0.09758 | 1.124 | 0.26110 |
| **previous_gr1** | -0.88237 | 0.21453 | -4.113 | 3.91e-05 *** |
| **poutcome_failure** | -1.28871 | 0.20852 | -6.180 | 6.40e-10 *** |
| **pdays_d1** | 0.81645 | 0.20713 | 3.942 | 8.09e-05 *** |
| **cons.price.idx_d1** | -0.89150 | 0.10308 | -8.648 | <2e-16 *** |
| **cons.conf.idx_bajo** | -0.51242 | 0.07566 | -6.773 | 1.26e-11 *** |

**Factor de varianza inflada (VIF)**

Revisamos entonces el factor de varianza inflada y se obtiene el siguiente resultado:

| **serit** | **emp.var.rate** | **marital_single** |
|---|---|---|
| 2.602167 | 2.501602 | 1.017366 |
| **contact_cellular** | **monthd_gr2** | **duration_d1** |
| 1.496718 | 1.553880 | 1.350700 |
| **campaign_d1** | **previous_gr1** | **poutcome_failure** |
| 1.019948 | 17.184850 | 10.808276 |
| **pdays_d1** | **cons.price.idx_d1** | cons.conf.idx_bajo |
| 8.004349 | 1.193678 | 1.814226 |

Según los valores obtenidos en la tabla anterior estamos frente a la existencia de multicolinealidad pues hay 2 variables que tienen un valor superior a 10 entonces procedemos eliminando la que tiene mayor valor (previous_gr1).

| | Estimate Std. | Error | z value | Pr(>—z—) |
|---|---|---|---|---|
| **(Intercept)** | 0.70670 | 0.11933 | 5.922 | 3.18e-09 *** |
| **serit** | -2.75627 | 0.09206 | -29.940 | <2e-16 *** |
| **emp.var.rate** | -0.34938 | 0.01960 | -17.822 | <2e-16 *** |
| **marital_single** | 0.14778 | 0.04633 | 3.190 | 0.00142 ** |
| **contact_cellular** | 1.11395 | 0.06328 | 17.604 | <2e-16 *** |
| **monthd_gr2** | 0.52090 | 0.05920 | 8.799 | <2e-16 *** |
| **duration_d1** | -3.07872 | 0.05265 | -58.480 | <2e-16 *** |
| **campaign_d1** | 0.11637 | 0.09762 | 1.192 | 0.23322 |
| **poutcome_failure** | -0.47767 | 0.06739 | -7.089 | 1.36e-12 *** |
| **pdays_d1** | 1.60553 | 0.07753 | 20.710 | <2e-16 *** |
| **cons.price.idx_d1** | -0.88953 | 0.10310 | -8.628 | <2e-16 *** |
| **cons.conf.idx_bajo** | -0.52194 | 0.07549 | -6.914 | 4.72e-12 *** |

**Factor de varianza inflada (VIF)**

Revisamos entonces el factor de influencia de varianza y se obtiene el siguiente resultado:

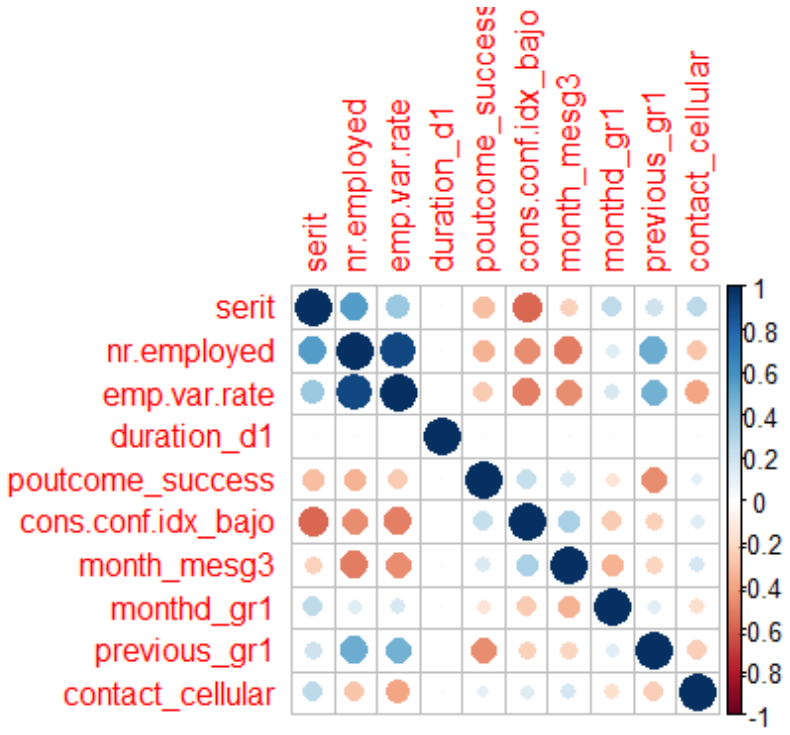



Figura 17: Gráfico de correlación de variables corregido

| **serit** | **emp.var.rate** | **marital_single** |
|---|---|---|
| 2.606264 | 2.494515 | 1.017335 |
| **contact_cellular** | **monthd_gr2** | **duration_d1** |
| 1.495721 | 1.554337 | 1.350502 |
| **campaign_d1** | **poutcome_failure** | **pdays_d1** |
| 1.019489 | 1.106640 | 1.137261 |
| **cons.price.idx_d1** | **cons.conf.idx_bajo** | |
| 1.193573 | 1.811409 | |

Con este proceso se ha corregido la multicolinealidad, por tanto, este es un buen modelo.

**Validación del modelo**

Usando el estadístico de prueba deviance con $p - value = 0$, lo que indica que es significativo.

### 7.2.7. Matriz de Confusión

La matriz de confusión es una herramienta muy útil para valorar cómo de bueno es un modelo de clasificación. Para evaluar nuestro modelo calcularémos el 'accuracy', como la proporicón entre las predicciones correctas que ha hecho el modelo y el total de predicciones.

$$Exactitud = \frac{\text{Predicciones correctas}}{\text{Número total de predicciones}}$$

Por otro lado, para profundizar un poco más y tener un cuenta los tipos de predicciones correctas e incorrectas que realiza el clasificador mostraremos la matriz de confusión.

| | 0 | 1 |
|---|---|---|
| 0 | 7126 | 180 |
| 1 | 639 | 286 |

De esta matriz de confusión podemos obtener la siguiente información:

$$Acurracy = 0{,}9004981$$

Lo que indica que un 90 % de la data clasifica correctamente.

**Observaciones**:

- La métrica accuracy (exactitud) no funciona bien cuando las clases están desbalanceadas.
- Acorde a lo visto en clase debemos balancer la data si la proporcion de 1 es menor al 5 % y dado que para nuestro caso $926/(7307+926) = 0,11$ es decir la proporcion es del 11 % no es necesario balancear la data. Por tanto del valor obtenido de Acurracy podemos decir que el porcentaje de la data clasificada correctamente es del 89 % por lo cual tenemos un buen modelo.

### 7.2.8. Tasa de error

La tasa de error definida como el cociente entre las predicciones incorrectas y el total de predicciones= 0,09004981 lo que nos indica que el porcentaje de la data clasificada incorrectamente es del 1 %, por lo cual tenemos un buen modelo.

### 7.2.9. Curva ROC

Es una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario según se varía el umbral de discriminación. Otra interpretación de este gráfico es la representación de la razón o ratio de verdaderos positivos (VPR = Razón de Verdaderos Positivos) frente a la razón o ratio de falsos positivos (FPR = Razón de Falsos Positivos) también según se varía el umbral de discriminación (valor a partir del cual decidimos que un caso es un positivo).
El área bajo la curva AUC puede tomar valores entre 0.5 y 1.

Para nuestro modelo el gráfico de la curva Roc es el siguiente:

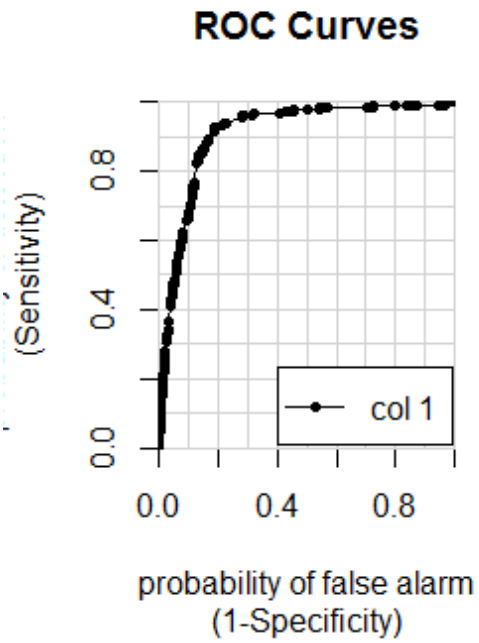



Figura 18: Curva ROC

En la figura anterior podemos notar que el área bajo bajo la curva está entre 1 y 0.9 por lo que se concluye que se obtuvo un buen modelo de predicción.

#### 7.2.10. Prueba de bondad de Ajuste de Hosmer-Lemeshow

La prueba de Hosmer-Lemeshow es una prueba estadística de bondad de ajuste para modelos de regresión logística . Se utiliza con frecuencia en modelos de predicción de riesgos. La prueba evalúa si las tasas de eventos observados coinciden o no con las tasas de eventos esperadas en subgrupos de la población del modelo. La prueba de Hosmer-Lemeshow identifica específicamente subgrupos como los deciles de los valores de riesgo ajustados. Los modelos para los que las tasas de eventos esperados y observados en los subgrupos son similares se denominan bien calibrados. Dada la explicación se establecen las hipótesis de la prueba son:

$$H_0 : \text{Las probabilidades observadas y esperadas son semejantes}$$

$$H_a : \text{Las probabilidades observadas y esperadas no son semejantes}$$

La prueba fue aplicada en R mediante la función hoslem.test del paquete 'generalhoslem'.
El p-valor obtenido es de 0.052 por lo que no se rechaza la hipótesis nula.

### 7.3. Modelo TTC

#### 7.3.1. Selección de Variables

Para este modelo se utilizaron solo variables que describen información particular del cliente, por consiguiente solo información idiosincrática. Como se mencionó en la sección 5, las variables idiosincráticas usadas usadas fueron: y, age, job, marital, education, housing, loan contact, monthd, day_of_week, serit .

### 7.3.2. Information Value (IV) y Estadístico KS

Se tomarán las variables con valores predictivos medios y fuertes, respectivamente.

| Variable | Test |
|---|---|
| serit | 1.283 |
| monthd | 1.077 |
| month | 0.407 |
| age | 0.256 |
| contact | 0.247 |
| job | 0.051 |
| marital | 0.028 |
| education | 0.018 |
| day_of_week | 0.005 |
| housing | 0.002 |
| loan | 0.000 |

Eliminamos así los predictores inútiles, que son: day_of_week, housing y loan.

### 7.3.3. Modelo LOGIT TTC

**MODELO INICIAL**

| | Estimate Std. | Error | z value | Pr(>—z—) |
|---|---|---|---|---|
| **(Intercept)** | -0.002826 | 0.407611 | -0.007 | 0.99447 |
| **age** | -0.001046 | 0.002113 | -0.495 | 0.62075 |
| **serit** | 1.745953 | 0.250073 | 6.982 | 2.91e-12 *** |
| **job_st.alto** | 0.051210 | 0.045965 | 1.114 | 0.26523 |
| **job_st.bajo** | 0.149200 | 0.053999 | 2.763 | 0.00573 ** |
| **job_st.medio** | NA | NA | NA | NA |
| **education_basica** | -0.030389 | 0.050486 | -0.602 | 0.54722 |
| **education_media** | -0.029234 | 0.050187 | -0.583 | 0.56023 |
| **education_superior** | NA | NA | NA | NA |
| **marital_divorced** | -0.458224 | 0.398607 | -1.150 | 0.25032 |
| **marital_married** | -0.482799 | 0.394841 | -1.223 | 0.22142 |
| **marital_single** | -0.378473 | 0.395440 | -0.957 | 0.33852 |
| **marital_unknown** | NA | NA | NA | NA |
| **contact_cellular** | 0.480688 | 0.062775 | 7.657 | 1.90e-14 *** |
| **contact_telephone** | NA | NA | NA | NA |
| **monthd_gr1** | -3.907513 | 0.231917 | -16.849 | <2e-16 *** |
| **monthd_gr2** | -3.442514 | 0.200125 | -17.202 | <2e-16 *** |
| **monthd_gr3** | NA | NA | NA | NA |
| **month_mesg1** | 0.054517 | 0.091303 | 0.597 | 0.55044 |
| **month_mesg2** | -0.597655 | 0.049755 | -12.012 | <2e-16 *** |
| **month_mesg3** | NA | NA | NA | NA |

Notamos que en el primer modelo presenta variables con valores NAs y procedemos a retirarlos, luego se retiran las variables que no son significativas en el siguiente orden;

- job_st.alto
- month_mesg1
- age
- marital_single
- education_basica
- education_media

**MODELO 1**

| | **Estimate Std.** | **Error** | **z value** | **Pr(>—z—)** |
|---|---|---|---|---|
| **(Intercept)** | -0.37857 | 0.06688 | -5.660 | 1.51e-08 *** |
| **serit** | 1.68327 | 0.22189 | 7.586 | 3.30e-14 *** |
| **job_st.bajo** | 0.11050 | 0.04577 | 2.414 | 0.01576 * |
| **marital_divorced** | -0.10309 | 0.06836 | -1.508 | 0.13154 |
| **marital_married** | -0.12492 | 0.04235 | -2.950 | 0.00318 ** |
| **contact_cellular** | 0.48599 | 0.06188 | 7.854 | 4.02e-15 *** |
| **monthd_gr1** | -3.82670 | 0.18411 | -20.785 | <2e-16 *** |
| **monthd_gr2** | -3.39435 | 0.17852 | -19.014 | <2e-16 *** |
| **month_mesg2** | -0.61326 | 0.04130 | -14.847 | <2e-16 *** |

El estadístico de prueba de deviance nos arroja con un pvalor= 0, por lo cual el modelo es significativo.

### 7.3.4. Multicolinealidad

Matriz de correlación:

| | **serit** | **job_st.bajo** | **marital_divorced** | **marital_married** |
|---|---|---|---|---|
| **serit** | 1.00000000 | -0.134299099 | 0.017591389 | 0.013784146 |
| **job_st.bajo** | -0.13429910 | 1.000000000 | 0.016787809 | 0.021509061 |
| **marital_divorced** | 0.01759139 | 0.016787809 | 1.000000000 | -0.435930642 |
| **marital_married** | 0.01378415 | 0.021509061 | -0.435930642 | 1.000000000 |
| **contact_cellular** | 0.26870910 | 0.007348522 | -0.003178444 | -0.064368235 |
| **monthd_gr1** | 0.25342178 | -0.080687208 | 0.014018997 | -0.012313525 |
| **monthd_gr2** | 0.31384494 | -0.024783152 | -0.002574123 | 0.060286831 |
| **month_mesg2** | 0.23020992 | 0.010495249 | 0.007828620 | 0.004147791 |

| | contact_cellular | monthd_gr1 | monthd_gr2 | month_mesg2 |
|---|---|---|---|---|
| **serit** | 0.268709101 | 0.25342178 | 0.313844935 | 0.230209919 |
| **job_st.bajo** | 0.007348522 | -0.08068721 | -0.024783152 | 0.010495249 |
| **marital_divorced** | -0.003178444 | 0.01401900 | -0.002574123 | 0.007828620 |
| **marital_married** | -0.064368235 | -0.01231352 | 0.060286831 | 0.004147791 |
| **contact_cellular** | 1.000000000 | -0.17847364 | 0.065650365 | 0.207844014 |
| **monthd_gr1** | -0.178473640 | 1.00000000 | -0.768622059 | -0.438540613 |
| **monthd_gr2** | 0.065650365 | -0.76862206 | 1.000000000 | 0.452582650 |
| **month_mesg2** | 0.207844014 | -0.43854061 | 0.452582650 | 1.000000000 |

Esta matriz de correlación nos indica que existe una alta correlación entre algunas variables lo que nos da un indicio de que existe multicolinealidad.

### 7.3.5. Factor inflación de varianza (VF)

Revisemos entonces el factor de influencia de varianza.

| serit | job_st.bajo | marital_divorced |
|---|---|---|
| 20.216818 | 1.044735 | 1.180907 |
| **marital_married** | **contact_cellular** | **monthd_gr1** |
| 1.191245 | 1.640524 | 19.859322 |
| **monthd_gr2** | **month_mesg2** | |
| 18.121710 | 1.145942 | |

Según los valores obtenidos hay 3 variables con valores superiores a 10 por tanto eliminamos la mas alta (monthd_gr1).

**MODELO 2**

| | Estimate Std. | Error | z value | Pr(>—z—) |
|---|---|---|---|---|
| **(Intercept)** | -1.17791 | 0.05918 | -19.903 | <2e-16 *** |
| **serit** | -2.81313 | 0.06117 | -45.991 | <2e-16 *** |
| **job_st.bajo** | 0.16187 | 0.04524 | 3.578 | 0.000346 *** |
| **marital_divorced** | -0.13825 | 0.06765 | -2.044 | 0.040985 * |
| **marital_married** | -0.18337 | 0.04193 | -4.373 | 1.23e-05 *** |
| **contact_cellular** | 1.50359 | 0.04807 | 31.280 | <2e-16 *** |
| **monthd_gr2** | 0.21021 | 0.05271 | 3.988 | 6.66e-05 *** |
| **month_mesg2** | -0.40052 | 0.04223 | -9.483 | <2e-16 *** |

Revisemos el factor de varianza inflada.

| **serit** | **job_st.bajo** | **marital_divorced** |
|---|---|---|
| 1.448695 | 1.039668 | 1.183262 |
| **marital_married** | **contact_cellular** | **monthd_gr2** |
| 1.191189 | 1.073214 | 1.595185 |
| **month_mesg2** | | |
| 1.227116 | | |

**Validación del modelo**

Notamos que ya no existe multicolinealidad. Usando el estadístico de prueba deviance nos muestra un $p - value = 0$, lo que indica que es significativo.

### 7.3.6. Matriz de Confusión

| | 0 | 1 |
|---|---|---|
| 0 | 7153 | 154 |
| 1 | 779 | 147 |

De esta matriz de confusión se obtiene la siguiente información:

$$Acurracy = 0{,}8866756$$

Lo que indica que un 88 % de la data clasifica correctamente.

### 7.3.7. Tasa de Error

$$Error = 0{,}1133244$$

Indica que el porcentaje de la data clasificada incorrectamente es del 1 %, por lo cual es un buen modelo.

### 7.3.8. Curva ROC

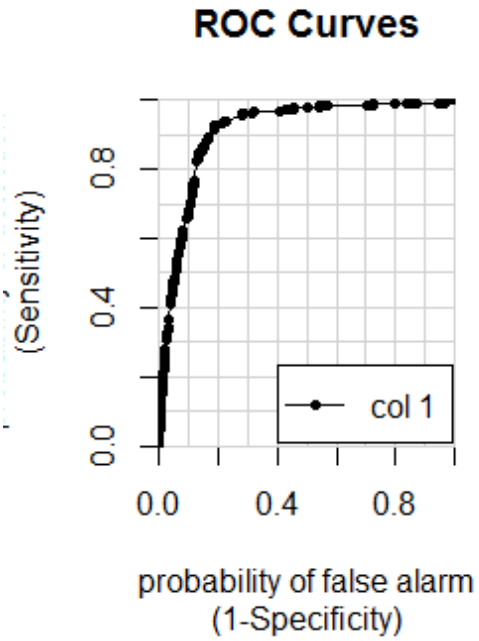



Figura 19: Curva ROC

La figura 19 podemos notar que el área bajo la curva es de 0,75 por lo cual obtuvimos un modelo aceptable de predicción.

## 7.4. Resultados

De los modelos obtenidos podemos concluir que el modelo PIT es mejor que el modelo TTC según los indicadores acuraccy 0.89 % y 0.75 % respectivamente , además de que el modelo PIT utiliza más variables que lo modelo TTC lo cual incide a ser un mejor modelo.

# 8. Creación de Grupos de Riesgo

Para la creación de los grupos homogéneos de riesgo se consideran los modelos PIT y TTC creados y validados en las secciones 7.2 y 7.3. Para la creación de estos grupos se usara la variable (prob.) obtenida de los modelos logisticos, a dicha variable se la separará en quintiles y se notará por las letras A,B,C,D y E a cada quintil, siendo la letra A correspondiente al primer quintil.Así, se analizará el número de ceros sobre el total en cada grupo y de esta manera se tendrá una visión macro de los grupos y se podrá analizar cuales grupos podrían unirse.

## 8.1. Grupos de Riesgo: Modelo PIT

Una vez realizado el análisis por grupos obtuvimos las siguiente tablas, la pimera hace referencia al número de ceros sobre el total y la segunda muestra las probabilidades a las que equivaldrían:

| mes | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | 3689 | 829 | 185 | 199 | 106 |
| 2 | 4036 | 1262 | 261 | 220 | 112 |
| 3 | 5507 | 1586 | 261 | 195 | 97 |
| 4 | 4855 | 678 | 225 | 131 | 112 |
| 5 | 63 | 0 | 5 | 0 | 0 |
| 6 | 2302 | 1233 | 187 | 202 | 108 |
| 7 | 2838 | 901 | 202 | 177 | 93 |
| 8 | 252 | 26 | 17 | 0 | 42 |
| 9 | 2302 | 54 | 89 | 0 | 2 |
| 10 | 2838 | 1389 | 302 | 177 | 153 |
| 11 | 662 | 105 | 0 | 0 | 0 |
| 12 | 158 | 62 | 0 | 0 | 0 |
| 13 | 725 | 80 | 0 | 0 | 0 |
| 14 | 252 | 58 | 0 | 0 | 0 |
| 15 | 410 | 18 | 44 | 41 | 0 |
| 16 | 315 | 47 | 63 | 0 | 0 |
| 17 | 158 | 14 | 55 | 44 | 0 |
| 18 | 221 | 65 | 49 | 46 | 13 |
| 19 | 158 | 14 | 44 | 39 | 6 |
| 20 | 189 | 14 | 19 | 0 | 6 |
| 21 | 189 | 7 | 16 | 0 | 12 |
| 22 | 284 | 14 | 18 | 0 | 12 |
| 23 | 221 | 14 | 19 | 0 | 0 |
| 24 | 95 | 202 | 36 | 4 | 20 |
| 25 | 32 | 158 | 52 | 18 | 25 |
| 26 | 32 | 101 | 33 | 17 | 25 |

| mes | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | 0,113 | 0,093 | 0,085 | 0,132 | 0,112 |
| 2 | 0,123 | 0,141 | 0,120 | 0,146 | 0,119 |
| 3 | 0,168 | 0,178 | 0,120 | 0,129 | 0,103 |
| 4 | 0,148 | 0,076 | 0,103 | 0,087 | 0,119 |
| 5 | 0,002 | 0,000 | 0,002 | 0,000 | 0,000 |
| 6 | 0,070 | 0,138 | 0,086 | 0,134 | 0,114 |
| 7 | 0,087 | 0,101 | 0,093 | 0,117 | 0,099 |
| 8 | 0,008 | 0,003 | 0,008 | 0,000 | 0,044 |
| 9 | 0,070 | 0,006 | 0,041 | 0,000 | 0,002 |
| 10 | 0,087 | 0,156 | 0,138 | 0,117 | 0,162 |
| 11 | 0,020 | 0,012 | 0,000 | 0,000 | 0,000 |
| 12 | 0,005 | 0,007 | 0,000 | 0,000 | 0,000 |
| 13 | 0,022 | 0,009 | 0,000 | 0,000 | 0,000 |
| 14 | 0,008 | 0,006 | 0,000 | 0,000 | 0,000 |
| 15 | 0,013 | 0,002 | 0,020 | 0,027 | 0,000 |
| 16 | 0,010 | 0,005 | 0,029 | 0,000 | 0,000 |
| 17 | 0,005 | 0,002 | 0,025 | 0,029 | 0,000 |
| 18 | 0,007 | 0,007 | 0,022 | 0,030 | 0,014 |
| 19 | 0,005 | 0,002 | 0,020 | 0,026 | 0,006 |
| 20 | 0,006 | 0,002 | 0,009 | 0,000 | 0,006 |
| 21 | 0,006 | 0,001 | 0,007 | 0,000 | 0,013 |
| 22 | 0,009 | 0,002 | 0,008 | 0,000 | 0,013 |
| 23 | 0,007 | 0,002 | 0,009 | 0,000 | 0,000 |
| 24 | 0,003 | 0,023 | 0,016 | 0,003 | 0,021 |
| 25 | 0,001 | 0,018 | 0,024 | 0,012 | 0,026 |
| 26 | 0,001 | 0,011 | 0,015 | 0,011 | 0,026 |

De donde obtenemos el siguiente gráfico de las probabilidades de cada grupo:

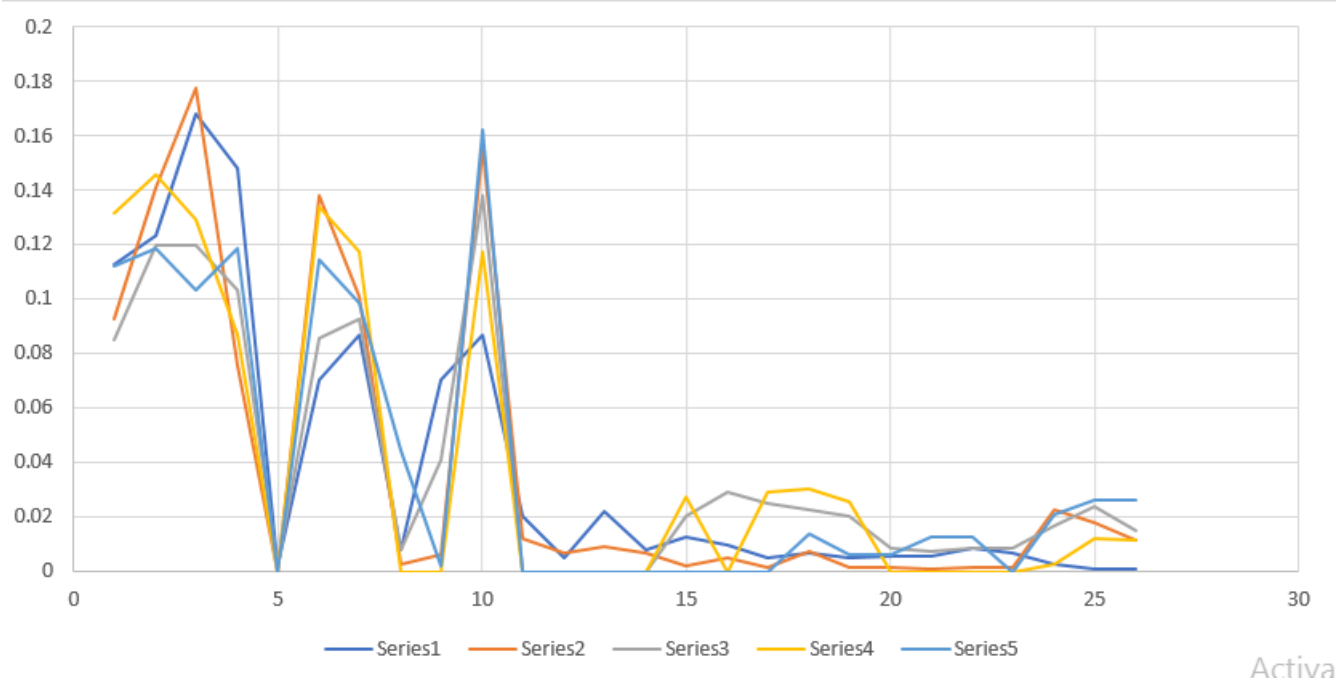



Figura 20

De este gráfico se puede observar que existe una acumulación de ceros, por lo cual se agruparon los meses según el siguiente criterio:

| mes | meses agrupados |
|---|---|
| 1 | 1 |
| 2 | 2 |
| 3 | 3 |
| 4 | 4 |
| 5 | 5,6,7,8 |
| 6 | 9 |
| 7 | 10 |
| 8 | 11-26 |

Dando como resultados las nuevas tablas del número de individuos en cada grupo y las probabilidades de cada grupo:

| mes | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | 3689 | 829 | 185 | 199 | 106 |
| 2 | 4036 | 1262 | 261 | 220 | 112 |
| 3 | 5507 | 1586 | 261 | 195 | 97 |
| 4 | 4855 | 678 | 225 | 131 | 112 |
| 5 | 3500 | 865 | 174 | 199 | 104 |
| 6 | 2302 | 1233 | 187 | 202 | 108 |
| 7 | 2838 | 901 | 202 | 177 | 93 |
| 8 | 4099 | 973 | 448 | 209 | 119 |

| mes | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | 0,117 | 0,23 | 0,22 | 0,56 | 0,5 |
| 2 | 0,128 | 0,35 | 0,31 | 0,62 | 0,53 |
| 3 | 0,175 | 0,44 | 0,31 | 0,55 | 0,46 |
| 4 | 0,154 | 0,188 | 0,267 | 0,37 | 0,53 |
| 5 | 0,111 | 0,24 | 0,206 | 0,56 | 0,49 |
| 6 | 0,073 | 0,342 | 0,222 | 0,57 | 0,51 |
| 7 | 0,09 | 0,25 | 0,24 | 0,499 | 0,44 |
| 8 | 0,13 | 0,27 | 0,532 | 0,59 | 0,56 |

Obteniendo como resultado el siguiente grafico de grupos:

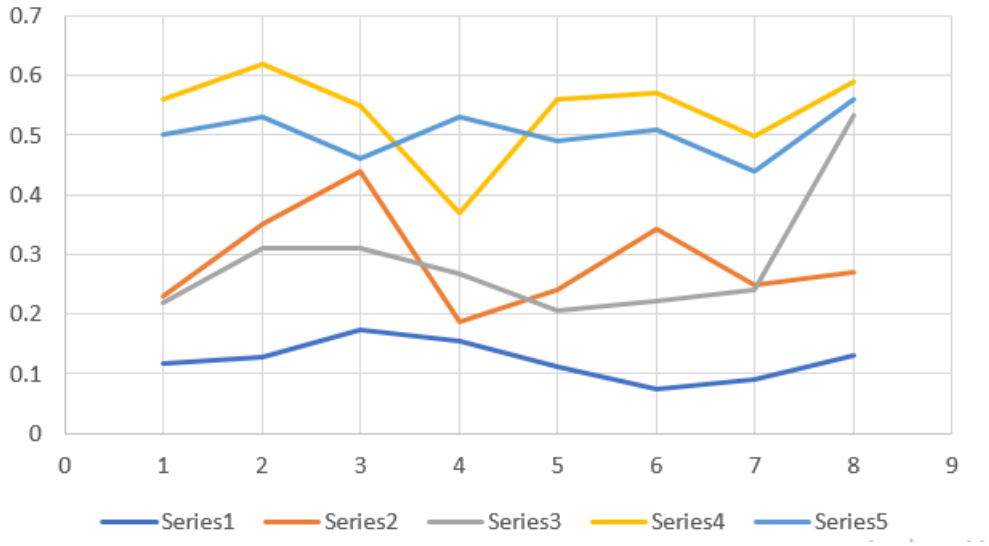



Figura 21

En el cual seria factible unir los quintiles 2 y 3 y los quintiles 4 y 5, con ello tendriamos las nuevas tablas:

| mes | A | BC | DE |
|---|---|---|---|
| 1 | 3689 | 1015 | 305 |
| 2 | 4036 | 1523 | 332 |
| 3 | 5507 | 1848 | 293 |
| 4 | 4855 | 903 | 244 |
| 5 | 3500 | 1039 | 303 |
| 6 | 2302 | 1420 | 310 |
| 7 | 2838 | 1104 | 270 |
| 8 | 4099 | 1422 | 328 |

| mes | A | BC | DE |
|---|---|---|---|
| 1 | 0,117 | 0,228 | 0,538 |
| 2 | 0,128 | 0,342 | 0,586 |
| 3 | 0,175 | 0,415 | 0,516 |
| 4 | 0,154 | 0,192 | 0,421 |
| 5 | 0,111 | 0,234 | 0,534 |
| 6 | 0,073 | 0,319 | 0,548 |
| 7 | 0,090 | 0,248 | 0,477 |
| 8 | 0,130 | 0,320 | 0,579 |

Y el gráfico final para los grupos homogéneos usando el modelo PIT sería:

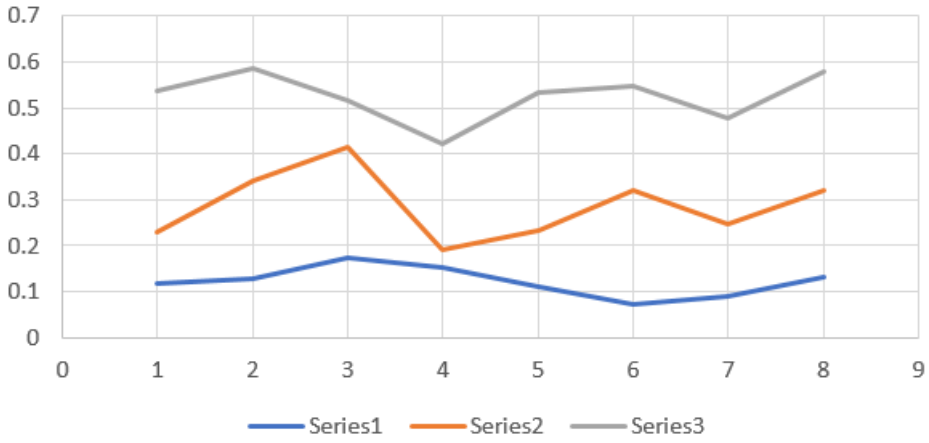



Figura 22

Donde:

- Series1 (grupo A): quintil 1.
- Series2 (grupo BC): quintil 2 y 3.
- Series3 (grupo DE): quintil 4 y 5.

Y podemos observar además que el grupo A toma valores entre [0,073 ; 0,175] , el grupo BC toma valores entre [0,192 ; 0,415] y el grupo DE toma valores entre [0,421 ; 0,586].

Finalmente presentamos un gráfico donde se muestra el porcentaje de cada grupo en cada mes:

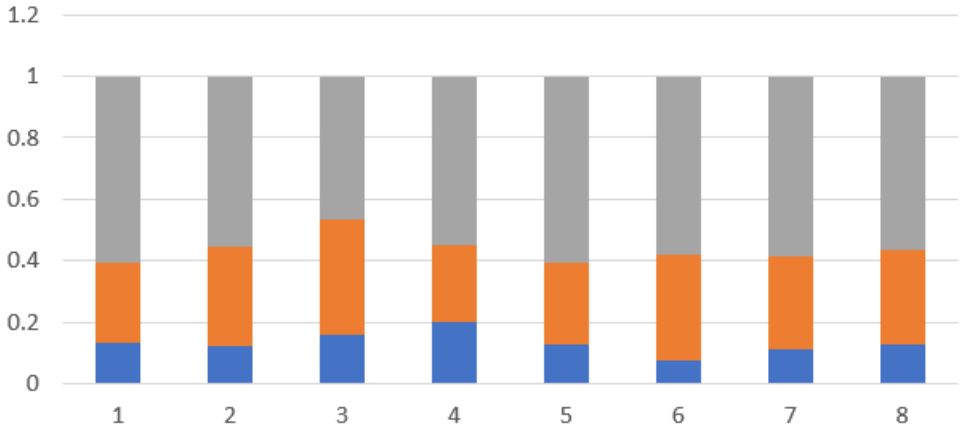



Figura 23

## 8.2. Grupos de Riesgo: Modelo TTC

Una vez realizado el análisis por grupos obtuvimos las siguiente tablas, la pimera hace referencia al número de ceros sobre el total y la segunda muestra las probabilidades a las que equivaldrían:

| mes | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | 4060 | 160 | 362 | 107 | 87 |
| 2 | 4367 | 147 | 371 | 109 | 83 |
| 3 | 6686 | 159 | 350 | 99 | 91 |
| 4 | 5185 | 162 | 377 | 91 | 89 |
| 5 | 68 | 0 | 0 | 0 | 0 |
| 6 | 3514 | 121 | 485 | 154 | 108 |
| 7 | 0 | 0 | 0 | 0 | 0 |
| 8 | 273 | 0 | 0 | 0 | 0 |
| 9 | 2456 | 161 | 350 | 70 | 86 |
| 10 | 3036 | 322 | 984 | 219 | 161 |
| 11 | 716 | 0 | 0 | 0 | 0 |
| 12 | 171 | 0 | 0 | 0 | 0 |
| 13 | 785 | 0 | 0 | 0 | 0 |
| 14 | 273 | 0 | 0 | 0 | 0 |
| 15 | 443 | 0 | 0 | 1 | 0 |
| 16 | 341 | 0 | 0 | 0 | 0 |
| 17 | 171 | 0 | 0 | 0 | 0 |
| 18 | 239 | 21 | 27 | 22 | 8 |
| 19 | 171 | 0 | 0 | 19 | 0 |
| 20 | 205 | 0 | 0 | 1 | 8 |
| 21 | 239 | 0 | 0 | 0 | 8 |
| 22 | 307 | 0 | 0 | 0 | 8 |
| 23 | 239 | 0 | 0 | 0 | 0 |
| 24 | 136 | 55 | 120 | 7 | 13 |
| 25 | 34 | 28 | 151 | 27 | 25 |
| 26 | 34 | 47 | 62 | 13 | 12 |

| mes | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | 0,119 | 0,116 | 0,100 | 0,114 | 0,111 |
| 2 | 0,128 | 0,106 | 0,102 | 0,116 | 0,105 |
| 3 | 0,196 | 0,115 | 0,096 | 0,106 | 0,115 |
| 4 | 0,152 | 0,117 | 0,104 | 0,097 | 0,113 |
| 5 | 0,002 | 0,000 | 0,000 | 0,000 | 0,000 |
| 6 | 0,103 | 0,088 | 0,133 | 0,164 | 0,137 |
| 7 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| 8 | 0,008 | 0,000 | 0,000 | 0,000 | 0,000 |
| 9 | 0,072 | 0,116 | 0,096 | 0,075 | 0,110 |
| 10 | 0,089 | 0,233 | 0,270 | 0,233 | 0,205 |
| 11 | 0,021 | 0,000 | 0,000 | 0,000 | 0,000 |
| 12 | 0,005 | 0,000 | 0,000 | 0,000 | 0,000 |
| 13 | 0,023 | 0,000 | 0,000 | 0,000 | 0,000 |
| 14 | 0,008 | 0,000 | 0,000 | 0,000 | 0,000 |
| 15 | 0,013 | 0,000 | 0,000 | 0,001 | 0,000 |
| 16 | 0,010 | 0,000 | 0,000 | 0,000 | 0,000 |
| 17 | 0,005 | 0,000 | 0,000 | 0,000 | 0,000 |
| 18 | 0,007 | 0,015 | 0,007 | 0,023 | 0,010 |
| 19 | 0,005 | 0,000 | 0,000 | 0,020 | 0,000 |
| 20 | 0,006 | 0,000 | 0,000 | 0,001 | 0,010 |
| 21 | 0,007 | 0,000 | 0,000 | 0,000 | 0,010 |
| 22 | 0,009 | 0,000 | 0,000 | 0,000 | 0,010 |
| 23 | 0,007 | 0,000 | 0,000 | 0,000 | 0,000 |
| 24 | 0,004 | 0,040 | 0,033 | 0,008 | 0,017 |
| 25 | 0,001 | 0,020 | 0,041 | 0,029 | 0,032 |
| 26 | 0,001 | 0,034 | 0,017 | 0,014 | 0,015 |

De donde obtenemos el siguiente gráfico de las probabilidades de cada grupo:

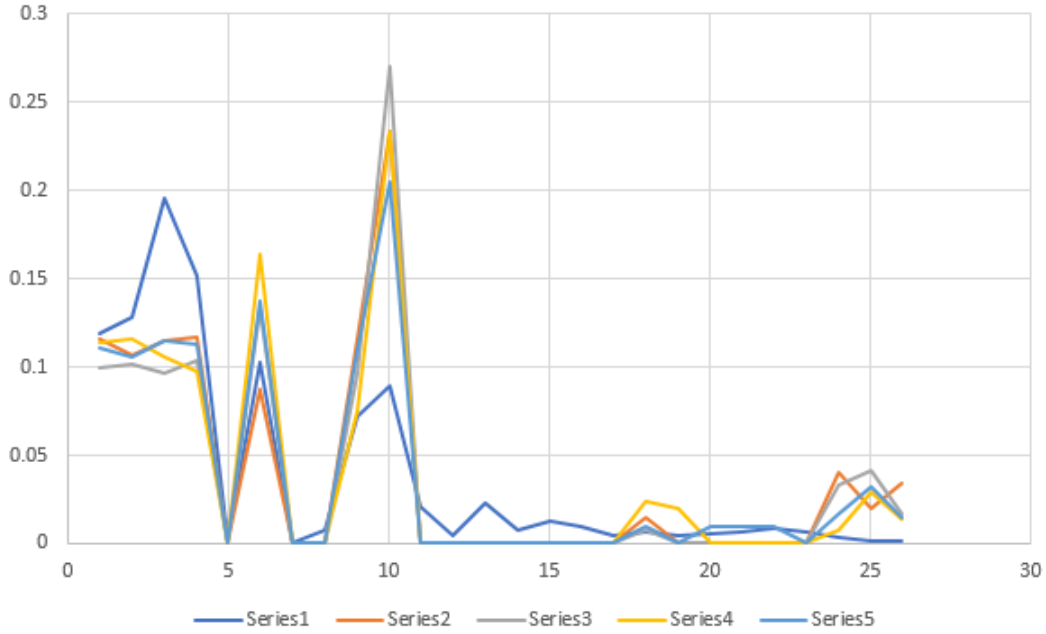



Figura 24

De este gráfico se puede observar que existe una acumulación de ceros, por lo cual fueron agrupados los meses según el siguiente criterio:

| mes | meses agrupados |
|---|---|
| 1 | 1 |
| 2 | 2 |
| 3 | 3 |
| 4 | 4 |
| 5 | 5,6,7,8 |
| 6 | 9 |
| 7 | 10 |
| 8 | 11-26 |

Dando como resultados las nuevas tablas del número de individuos en cada grupo y las probabilidades de cada grupo:

| mes | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | 4060 | 160 | 362 | 107 | 87 |
| 2 | 4367 | 147 | 371 | 109 | 83 |
| 3 | 6686 | 159 | 350 | 99 | 91 |
| 4 | 5185 | 162 | 377 | 91 | 89 |
| 5 | 3855 | 121 | 485 | 154 | 108 |
| 6 | 2456 | 161 | 350 | 70 | 86 |
| 7 | 7888 | 322 | 984 | 219 | 161 |
| 8 | 4503 | 151 | 360 | 90 | 82 |

| mes | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | 0,119 | 0,257 | 0,176 | 0,424 | 0,428 |
| 2 | 0,128 | 0,261 | 0,180 | 0,402 | 0,392 |
| 3 | 0,196 | 0,239 | 0,170 | 0,440 | 0,423 |
| 4 | 0,152 | 0,220 | 0,183 | 0,432 | 0,432 |
| 5 | 0,113 | 0,370 | 0,235 | 0,523 | 0,322 |
| 6 | 0,072 | 0,169 | 0,170 | 0,419 | 0,429 |
| 7 | 0,231 | 0,527 | 0,478 | 0,783 | 0,859 |
| 8 | 0,132 | 0,216 | 0,175 | 0,396 | 0,402 |

Obteniendo como resultado el siguiente gráfico de grupos:

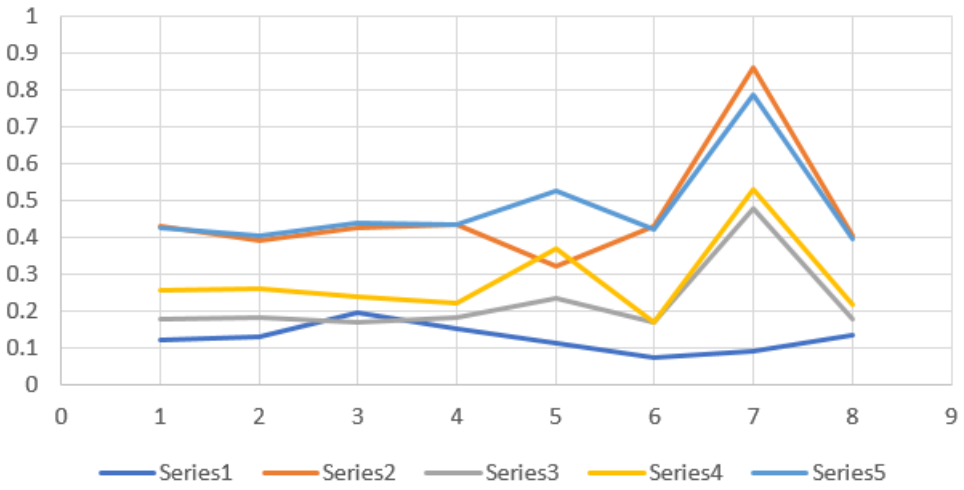



Figura 25

En el cual seria factible unir los quintiles 2 y 3 y los quintiles 4 y 5, con ello tendriamos las nuevas tablas:

| mes | A | BC | DE |
|---|---|---|---|
| 1 | 4059,6 | 522 | 194 |
| 2 | 4366,6 | 518 | 191 |
| 3 | 6686,3 | 508 | 190 |
| 4 | 5185,3 | 539 | 180 |
| 5 | 3854,9 | 605 | 262 |
| 6 | 2456,2 | 510 | 157 |
| 7 | 3036,1 | 1306 | 380 |
| 8 | 4503 | 511 | 171 |

| mes | A | BC | DE |
|---|---|---|---|
| 1 | 0,119 | 0,215 | 0,312 |
| 2 | 0,128 | 0,213 | 0,308 |
| 3 | 0,196 | 0,209 | 0,305 |
| 4 | 0,152 | 0,222 | 0,290 |
| 5 | 0,113 | 0,249 | 0,421 |
| 6 | 0,072 | 0,210 | 0,252 |
| 7 | 0,231 | 0,537 | 0,612 |
| 8 | 0,132 | 0,210 | 0,276 |

Y el gráfico final para los grupos homogéneos usando el modelo TTC sería:

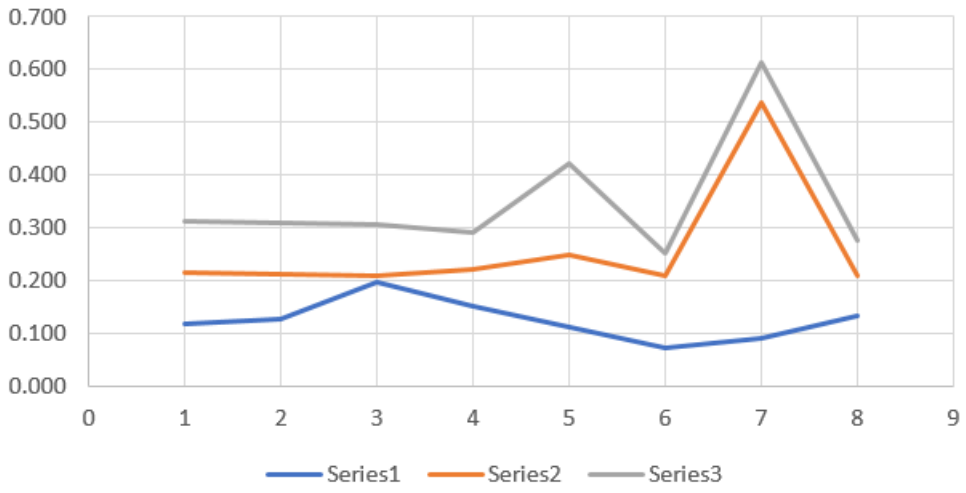



Figura 26

Donde:

- Series1 (grupo A): quintil 1.
- Series2 (grupo BC): quintil 2 y 3.
- Series3 (grupo DE): quintil 4 y 5.

Finalmente presentamos un gráfico donde se muestra el porcentaje de cada grupo en cada mes:

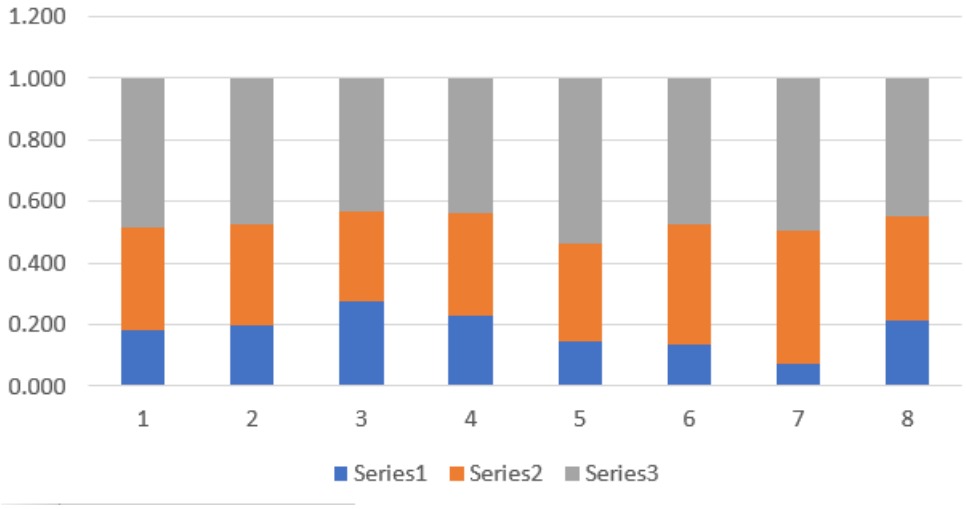



Figura 27

## 8.3. Validación de los grupos de riesgo

### 8.3.1. Prueba de Dunnet T3

**Modelo PIT**

Después de haber obtenido la clasificación de los grupos, vamos a proceder con la validación para el modelo PIT, pues se debe verificar que la tasa de mora de un grupo homogéneo es diferente a la tasa de mora de otro. Utilizando la prueba de Dunnett T3, donde:

$H_0$ : Las medias de los grupos son iguales..

$H_a$ : Las medias de los grupos son diferentes.

Se obtuvieron los siguientes resultados:

| | A | BC |
|---|---|---|
| BC | 2e-16 | |
| DE | 8.3e-12 | 2e-16 |

Podemos concluir que los grupos de riesgo del modelo PIT son válidos, puesto que al usar el test de Dunnett T3, de éste se obtiene un $pval < 0{,}05$, lo que implica que se rechaza la hipótesis nula, es decir, las medias de los grupos son diferentes.

**Modelo TTC**

Realizando la prueba Dunnett T3 con los grupos del modelo TTC se obtuvieron los siguientes resultados:

| | A | BC |
|---|---|---|
| BC | 1.0e-08 | |
| DE | 0.0035 | 0.0099 |

Podemos concluir que los grupos de riesgo del modelo PIT son válidos, puesto que al usar el test de Dunnett T3, de éste se obtiene un $pval < 0{,}05$, lo que implica que se rechaza la hipótesis nula, es decir, las medias de los grupos son diferentes.

# 9. Alocación del capital de la cartera crediticia

## 9.1. Pérdida Esperada

La pérdida esperada se define como el valor esperado de pérdida por riesgo crediticio en un horizonte de tiempo determinado.Basado en 3 importantes aspectos que son

- Probabilidad de incumplimiento (PD).
- Exposición al momento de incumplimiento (EAD).
- Pérdida en caso de incumplimiento (LGD).

Dados estos parámetros, se obtiene el valor de la pérdida esperada de la siguiente manera:

$$PE = PD * EAD * LGD$$

**Probabilidad de incumplimiento (PD):** La probabilidad de impago es la probabilidad prevista para que un prestatario se declare insolvente y deje de pagar sus cuotas de amortización.

**Exposición al momento de incumplimiento (EAD):** Es el tamaño de la deuda, es la cantidad pendiente o saldo al que ocurre el incumplimiento.

**Pérdida en caso de incumplimiento (LGD):** Es una estimación de la parte que realmente se pierde en caso de incumplimiento del cliente.

**VAR:** El VAR es la medición de la máxima pérdida esperada dado un horizonte de tiempo.

## 9.2. Pérdida Esperada PIT

### 9.2.1. PD

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:

| | Grupo A | Grupo B | Grupo C |
|---|---|---|---|
| PD | 0.122 | 0.287 | 0.525 |

### 9.2.2. EAD

Por otro lado, el modelamiento para obtener el valor del EAD de cada grupo es un trabajo complejo, por lo cual, se han tomado los siguientes valores para el EAD, cumpliendo la condición de grupos en donde,

$$EAD_A > EAD_B > EAD_C$$

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.

| | Grupo A | Grupo B | Grupo C |
|---|---|---|---|
| EAD | 10000 | 5000 | 3000 |

### 9.2.3. LGD

En esta parte, para el cálculo del LGD se ha realizado por grupo mediante una simulación de una distribución Beta de parámetros $(0{,}5; 0{,}5)$, teniendo como resultado la siguiente gráfica:

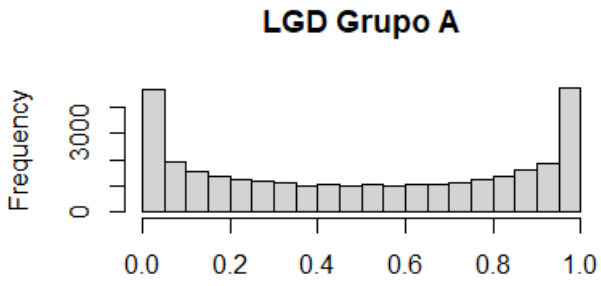



Figura 28

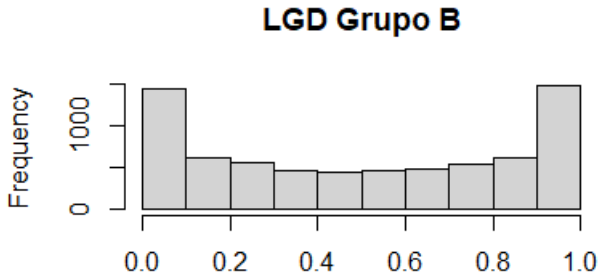



Figura 29

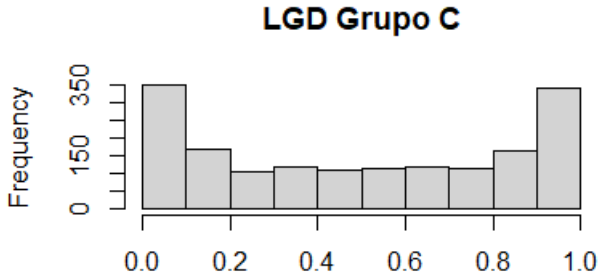



Figura 30

Después de la simulación del LGD para cada grupo, se obtiene su media verificando la condición:

$$LGD_C > LGD_B > LGD_A$$

Por lo tanto, el LGD para cada grupo es:

| | A | B | C |
|---|---|---|---|
| LGD | 0.47 | 0.48 | 0.51 |

### 9.2.4. Distribución pérdida

Conociendo los parámetros para el cálculo de la pérdida esperada, se grafica la distribución de la misma.

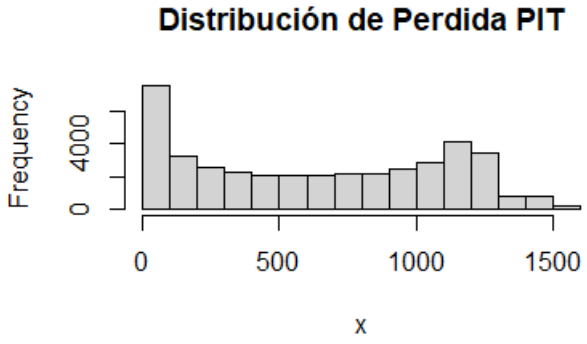



Figura 31

Ahora dada la gráfica podemos obtener los valores para la cartera como son:

| | |
|---|---|
| PE cartera | 639.45 |
| VAR | 1436.33 |
| CE | 796,88 |

Realizando un análisis similar por meses tenemos que los valores que debería guardar el banco en cada mes:

| Mes | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| CE | 878.9 | 740.7 | 761.8 | 729.5 | 784.6 | 737.4 | 865.2 | 805.1 |

Por lo que se puede observar en la siguiente gráfica los valores del capital económico para cada mes los cuales varian notablemente:

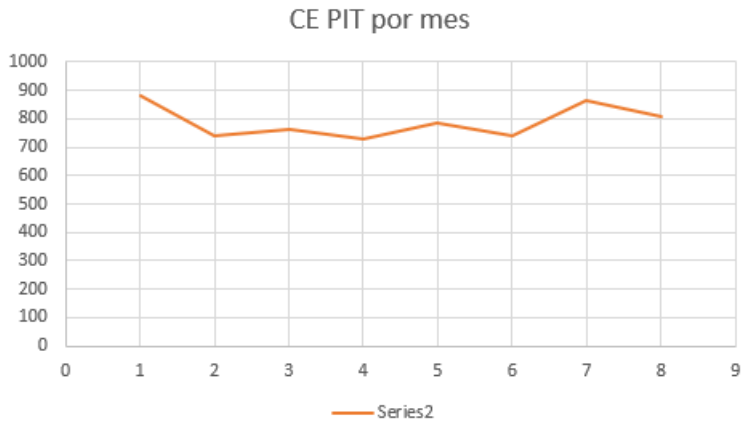



Figura 32

## 9.3. Pérdida Esperada TTC

### 9.3.1. PD

De igual forma una vez realizados lo grupos homogéneos podemos recuperar de ellos las PD para cada grupo, hemos obtenido los valores siguientes:

| | Grupo A | Grupo B | Grupo C |
|---|---|---|---|
| PD | 0.125 | 0.258 | 0.347 |

### 9.3.2. EAD

Se han tomado los siguientes valores para el EAD: El modelamiento para la obtención del valor del EAD de cada grupo se tomaron los siguientes valores cumpliendo la condición de grupos en donde;

$$EAD_A > EAD_B > EAD_C$$

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.

| | Grupo A | Grupo B | Grupo C |
|---|---|---|---|
| EAD | 10000 | 5000 | 3000 |

### 9.3.3. LGD

Para el cálculo del LGD se ha realizado por grupo la simulación de una distribución Beta de parámetros $(0{,}5; 0{,}5)$, teniendo como resultado la siguiente gráfica:

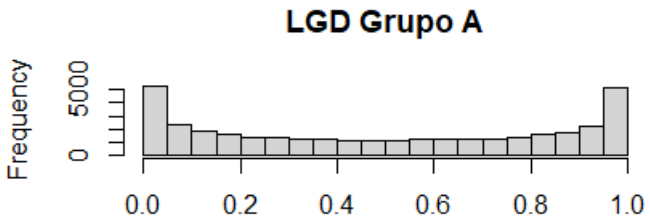



Figura 33

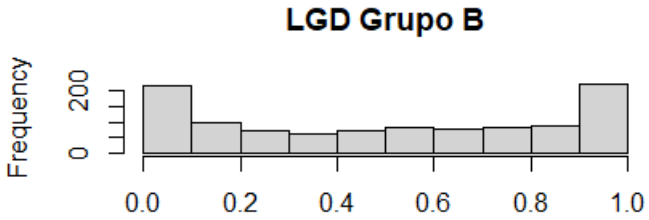



Figura 34

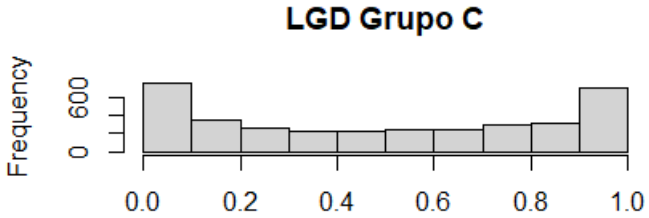



Figura 35

Después de la simulación del LGD para cada grupo, se obtiene su media verificando la condición:

$$LGD_C > LGD_B > LGD_A$$

Por lo tanto, el LGD para cada grupo es:

| | A | B | C |
|---|---|---|---|
| LGD | 0.47 | 0.47 | 0.48 |

### 9.3.4. Distribución pérdida

Conociendo los parámetros para el cálculo de la pérdida esperada, se grafica la distribución de la misma.

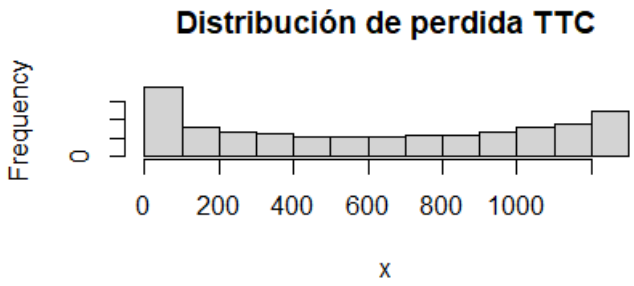



Figura 36

Dada la gráfica anterior podemos obtener los valores para la cartera como son:

| | |
|---|---|
| PE cartera | 618.74 |
| VAR | 1249.81 |
| CE | 631.06 |

Realizando un analisis similar por meses tenemos que los valores que deberia guardar el banco en cada mes serían:

| Mes | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| CE | 642.4 | 625.4 | 615.7 | 613.0 | 630.8 | 623.9 | 676.6 | 624.4 |

Por lo que se puede observar en la siguiente gráfica los valores del capital económico para cada mes los cuales varian menos que la metodología PIT:

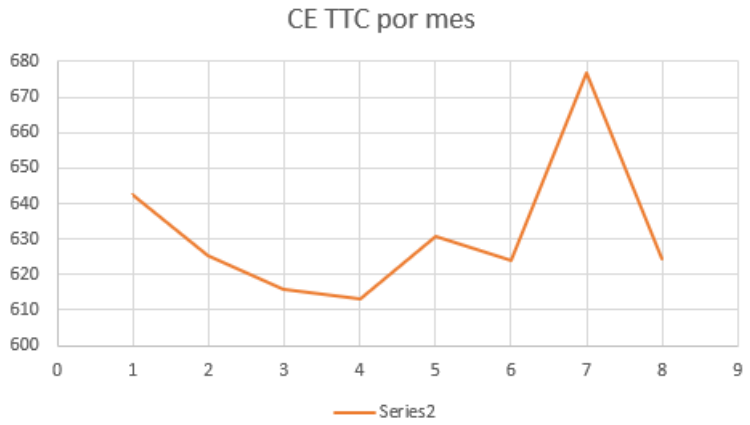



Figura 37

## 10. Conclusiones

- Nuestro modelo de Scoring creado nos ayudó a clasificar a aquellos que solicitan crédito en tipos de riesgos 'malo' y 'bueno'.

- El modelo PIT tiende a seguir una linea recta a diferencia del modelo TTC que tiende a seguir el ciclo económico.
- La clasificación de grupos homogéneos que han sido definidos por su calificación crediticia para el modelo PIT se obtuvieron 3 grupos bien identificados. Por otra parte, en este informe para el método TTC, se trató de identificar los grupos pero vale mencionar que se obtuvieron 3 de los cuales en 2 se puede observar fácilmente siguen el ciclo económico pero el último a pesar de ser distinto no se puede observar de manera fácil la tendencia a seguir el ciclo económico.
- Bajo la filosofía TTC, se observó que el capital económico para todos los meses fue similar, se concluyé que esto se da debido a que solo intervienen factores idiosincráticos. Por otro lado, en la filosofía PIT si es notable la variación de capital, puesto que al intervenir las variables macroeconómicas como lo son los indicadores diarios influyen para que la entidad financiera construya su capital económico y evitar pérdidas por incumplimiento de pago.

# 11. Referencias


- Jeffrey M. Wooldridge Introductory Econometrics A Modern Approach (2012)
- ES The Credit Scoring Toolkit Theory
- The Credit Scoring Toolkit Theory and Practice for Retail Credit Risk Management and Decision Automation (2007)
- Gestion de Riesgos en Entidades Financieras- Gerencia de Riesgos
- Modelos de Gestión del Riesgo de Crédito (Oscar Basso, 2013).
- An Introduction to Analysis of Financial Data with R (Ruey S. Tsay, 2013)
- http://www.scielo.org.pe/scielo.php?script=sci_arttextpid=S2077-18862010000100005: :text=Concepto
- https://connectamericas.com/es/content/lo-que-usted-debe-saber-sobre-el-credit-scoring
- http://jotvirtual.ucoz.es/RIESGOFINANC/5-VALOR_EN_RIESGO.pdf
- https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc