Universidad Francisco Marroquín
Catedrático: María Isabel Avila Rigalt
Auxiliar: Wilder de Jesús Villeda
Data Science for Finance

“INFORME CASO #1”

Estuardo García - 20220112

Índice

Introducción

Este proyecto tuvo como objetivo aplicar un modelo de regresión multivariada para predecir la cantidad de clientes activos mensuales de una institución financiera. El análisis se basó en bases de datos históricos que incluyeron tasas de interés, saldos promedio mensuales y variables macroeconómicas como (tasas lider, ingreso por divisas, exportaciones fob y inflación). Se aplicaron técnicas de ajuste de distribuciones, simulaciones estocásticas y pruebas de supuestos para garantizar la validez del modelo bajo el marco teórico de Gauss-Markov.

Exploración de Datos y Ajuste de Distribuciones

Se inició el análisis con la exploración estadística de las tasas activas mensuales y los saldos promedio por cliente. Utilizando funciones gráficas y el paquete fitdistrplus, se ajustaron distintas distribuciones teóricas. Los resultados mostraron que la distribución gamma ofrecía el mejor ajuste en ambos casos, según el criterio de Akaike. A partir de estos ajustes se generaron 10,000 simulaciones de tasas y saldos.

Estas simulaciones se utilizaron para calcular ingresos mensuales esperados por interés, multiplicando tasas simuladas por saldos simulados. Para representar un caso realista, se integraron las proyecciones de cantidad de clientes para los primeros tres meses, generando un archivo llamado intereses.csv con los ingresos proyectados para ese período.

Escenarios de Ingreso y Evaluación de Hipótesis

Se estimaron escenarios de ingresos bajo distintos niveles de confianza. En concreto, se calcularon los percentiles 1%, 5%, 95% y 99% para definir los escenarios extremos de menor y mayor ingreso mensual. Estos intervalos sirvieron también para evaluar la validez de una afirmación del Gerente de Banca Personas, quien estimaba un saldo promedio de Q13,000 por cliente.

Al comparar dicho valor con el rango comprendido entre los percentiles 5% y 95% de los saldos simulados, se concluyó que la afirmación era estadísticamente realista. Esto demuestra cómo la simulación permite contrastar escenarios con base empírica.

Construcción del Modelo de Regresión

La siguiente fase consistió en la construcción de un modelo de regresión lineal multivariada. Se separaron los datos en un conjunto de entrenamiento (meses 1 al 36) y otro de prueba (meses 37 al 48). El primer modelo incluyó todas las variables: proporción, saldo promedio, consumos recientes, tasa líder, ingresos por divisas, exportaciones e inflación.

Tras revisar el resumen estadístico del modelo y los niveles de significancia, se identificó que las variables más relevantes eran "saldo promedio" y "tasa líder". Con base en esto, se construyó un modelo optimizado que mantuvo un R^2 ajustado de 94.84%, reduciendo la complejidad sin perder poder predictivo.

Validación de Supuestos del Modelo

Se aplicaron pruebas formales para validar los supuestos del modelo bajo el teorema de Gauss-Markov:

- **Homoscedasticidad**: El gráfico de residuos contra valores ajustados mostró dispersión constante. Esto se confirmó con el test de Breusch-Pagan ($p = 0.4173$), que no rechazó la hipótesis de varianza constante.
- **Normalidad de residuos**: El test de Shapiro-Wilk ($p = 0.8935$) indicó que los residuos seguían una distribución normal.
- **Independencia de errores**: El test de Durbin-Watson ($DW = 1.4108$, $p = 0.01511$) confirmó que no existía autocorrelación positiva.

Estos resultados positivos avalan la solidez teórica del modelo.

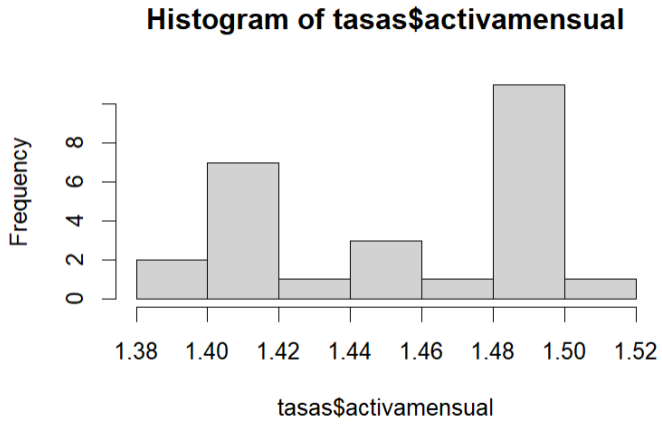
Evaluación del Modelo con Datos de Prueba

El modelo optimizado se evaluó sobre el conjunto de prueba. Se obtuvo un error absoluto medio (MAE) de 1826.44 y una precisión (accuracy) del 94.86%.

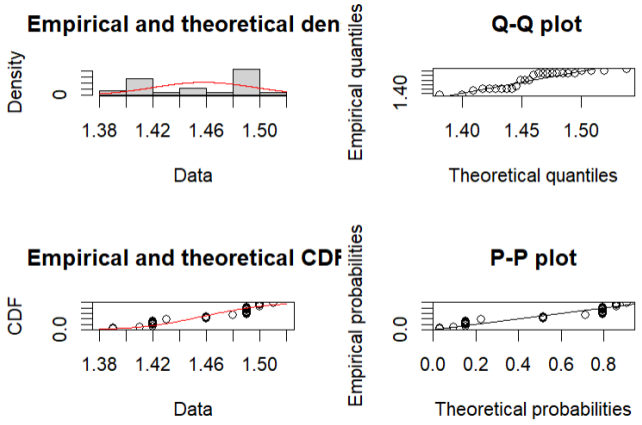
Conclusión

Se desarrolló un modelo de regresión multivariada simple, explicativo y altamente preciso, validado tanto empírica como teóricamente. Cumple con los supuestos del modelo lineal clásico y demuestra una capacidad predictiva muy elevada. Este modelo representa una herramienta robusta para la proyección de clientes activos en contextos financieros mensuales.

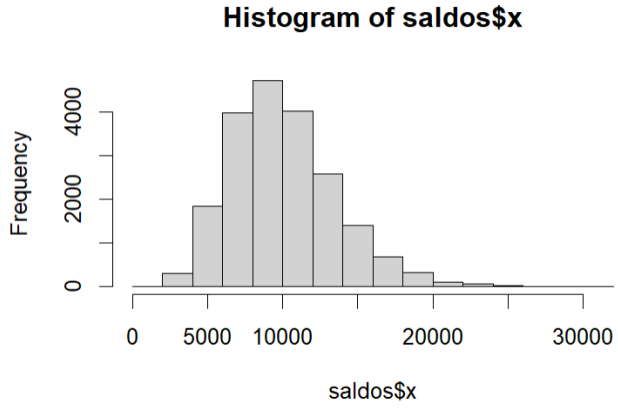
Histogramas



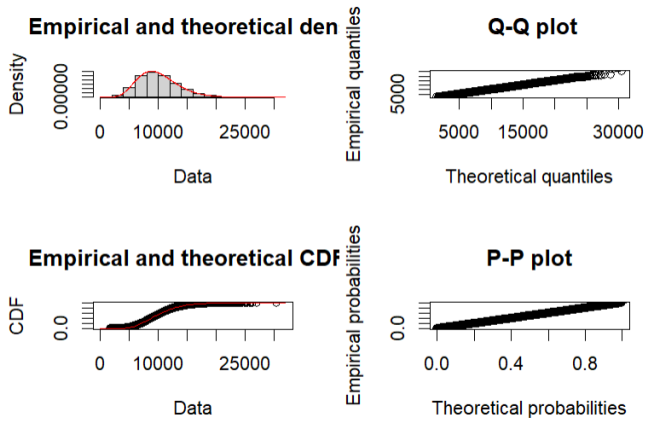
-Tasas Activa mensual



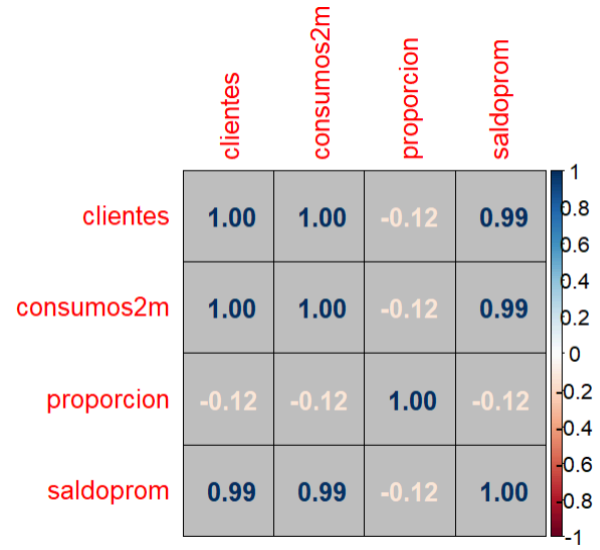
-Tasa Activa mensual (Gamma)



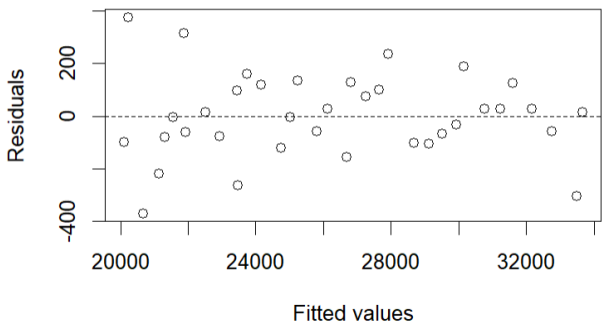
-Saldos



-Saldos (Gamma)



-Mapa de Calor



-Gauss Markov

```
        studentized Breusch-Pagan test

data:  regresionmultivariada2
BP = 5.1499, df = 2, p-value = 0.07616
```

-Breusch Pagan Test

```
        Shapiro-Wilk normality test

data:  resid(regresionmultivariada2)
W = 0.98543, p-value = 0.9071
```

-Shapiro Wilk

```
       Durbin-Watson test

data:  regresionmultivariada2
DW = 2.3169, p-value = 0.7288
alternative hypothesis: true autocorrelation is greater than 0
```

-Durbin Watson

```
Residuals:
    Min      1Q  Median      3Q     Max
-222.23 -104.49   -1.70   69.15  320.16

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      481.76752  546.02120   0.882   0.3851
proporcion         0.51582    3.44826   0.150   0.8822
saldoprom          0.05763    0.06705   0.859   0.3974
consumos2m         0.82605    0.03712  22.254   <2e-16 ***
tasaslider       127.18087   60.56243   2.100   0.0449 *
ingresopordivisas  0.37658    0.24576   1.532   0.1367
exportaciones      0.18043    0.29779   0.606   0.5495
inflacion         -0.65051   13.25909  -0.049   0.9612
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 159.3 on 28 degrees of freedom
Multiple R-squared:  0.9988,    Adjusted R-squared:  0.9985
F-statistic:  3325 on 7 and 28 DF,  p-value: < 2.2e-16
```

-Modelo 1 (Sin modificaciones)

```
Residuals:
    Min      1Q  Median      3Q     Max
-370.92  -85.78    4.77  104.23  375.55

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 582.04783  344.44085   1.690     0.10
consumos2m    0.88051    0.01699  51.821   <2e-16 ***
tasaslider   86.53746   54.22983   1.596     0.12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 165 on 33 degrees of freedom
Multiple R-squared:  0.9985,    Adjusted R-squared:  0.9984
F-statistic: 1.085e+04 on 2 and 33 DF,  p-value: < 2.2e-16
```

-Modelo 2 (Modificado)