# MODELO DE REGRESIÓN LOGÍSTICA PARA ESTIMAR LA PROBABILIDAD DE RETENCIÓN DE CLIENTES PROFESIONALES, EN UNA EMPRESA MINORISTA

*Logistic regression model to estimate the probability of retention of professional customers in a retail company.*

**Pedro Pablo Morales Ortiz**
Mtr. en Estadística Aplicada
pepamoralesortiz@gmail.com

**José Rolando Chávez Salazar**
Mtr. en Administración Industrial
rolando.chavez@gmail.com

Recibido: 24 de noviembre de 2023 | Revisado: 3 de mayo de 2024 | Aprobado: 10 de junio de 2024

## Resumen

La empresa requiere enfocar sus esfuerzos de mercadeo en desarrollar la lealtad en el segmento de clientes profesionales. Por ende, necesita contar con un sistema de información que indique la propensión a perder a sus clientes.

El estudio busca aplicar un modelo de regresión logística para predecir la retención de clientes profesionales. Se analizan variables que miden el comportamiento del cliente, como las variables RFM, y otras segmentaciones que se obtienen mediante el análisis de la base de datos transaccional de la empresa.

Se presenta el análisis de segmentación de los clientes y el modelo mejor ajustado a los datos reales, con una exactitud de predicción de 69.39%.

Se destaca la utilidad del modelo de regresión como herramienta para la toma de decisiones estratégicas y se recomienda a la empresa utilizarlo para mejorar sus estrategias de retención y fidelización de clientes.

## Palabras clave

Modelo de regresión, K medias, correlación, retención de clientes.

## *Abstract*

*The company needs to focus its marketing efforts on developing loyalty in the professional customer segment. Therefore, it needs to have an information system that indicates the propensity to lose its customers.*

*This study aims to build a logistic regression model to predict the retention or churn of professional customers. For this purpose, variables that measure customer behavior, such as RFM variables, and other segmentations obtained through the analysis and data mining of the company's transactional database are analyzed.*

*The segmentation analysis of the customers and the best-fitting logistic model to the actual data are presented, with a prediction accuracy of 69.39%.*

*The conclusion highlights the utility of the model as a tool for strategic decision-making and recommends that the company use it to improve its customer retention and loyalty strategies.*

## *Keywords*

*RFM, K means, correlation, customer churn, probability.*

## Introducción

El estudio analiza a los clientes profesionales de la empresa a partir de bases de datos transaccionales para identificar el modelo que permite inferir la probabilidad de retención o pérdida de estos.

Estudios previos estiman la probabilidad de compra basándose en las variables RFM (Aleksandrova, 2018). La segmentación de los clientes toma una fuerte importancia en este análisis, donde Cuadros, et al. (2017) y Dogan et al. (2018), lo resuelven usando un enfoque en análisis de datos multivariados.

En el presente estudio, se analizan las variables del método RFM obtenidas por minería de datos. Dichas variables son convertidas en variables categóricas utilizando el método de percentiles y el método de agrupación por K medias. Posteriormente, se realiza un análisis de correlación para determinar las variables a utilizar dentro del análisis de regresión.

Finalmente, se determina un modelo con una exactitud de predicción de 69.39% de los casos; pero el riesgo de basar el servicio en dicho modelo de predicción es únicamente de 12.53%.

## Desarrollo del estudio

Se realiza un análisis transversal de los datos transaccionales de 2021 y 2022. Los primeros 18 meses se utilizan para obtener las variables regresoras, y los 6 meses más recientes para determinar la retención o pérdida del cliente.

La siguiente fase corresponde al análisis de las variables reciencia (R), frecuencia (F) y monto promedio (M) para hacer su segmentación por medio del método de percentiles y el algoritmo de agrupación por K medias. Se analiza también la correlación entre las variables regresoras y la variable de resultado utilizando pruebas de independencia $X^2$.

Por último, se construyen múltiples modelos de regresión logística con diferentes combinaciones de variables.

Los valores estimados se convierten en una variable dicotómica asumiendo que un valor superior a 50% indica la retención del cliente y caso contrario indica su pérdida. Con esto, se comparan los valores estimados y los valores reales mediante matrices de confusión (Ohsaki et al., 2017) para elegir el modelo con mayor exactitud.

## Resultados obtenidos

Inicialmente, se analizan las variables RFM para su categorización usando dos métodos: el método de percentiles y una segmentación por K medias. De este último, se obtienen 3 segmentos suficientemente heterogéneos.

**Tabla 1.**

*Segmentos A, B y C obtenidos por algoritmo de K*

| | n | F | M | R | TR% |
|---|---|---|---|---|---|
| B | 555 | 57.9 | 5,743.50 | 12.5 | 76.6% |
| C | 1,069 | 138.4 | 4,272.90 | 167.1 | 37.0% |
| A | 914 | 17.0 | 3.769.90 | 199.5 | 36.2% |

*Nota.* Cantidad de clientes (n), frecuencia promedio en días (F), montopromedio quetzales (M), reciencia promedio en días (R) y tasa de retención (TR%). Elaboración propia

Asimismo, se obtienen agrupaciones percentiles de las variables RFM con sus respectivas interacciones (RF, RM, M).

El análisis de correlación, indica una relación significativa para las variables descritas, exceptuando el canal de ventas ($p = 0.43$), y una relación débil para la condición de pago ($p = 0.02$) y la agrupación de F percentil ($p = 0.02$).

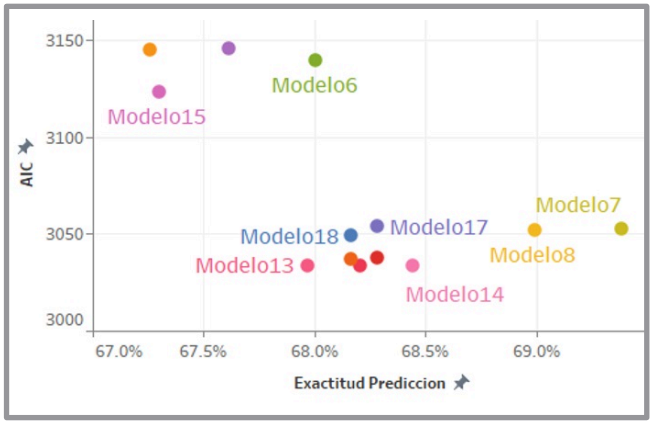
Para la construcción de modelos de regresión se iteran múltiples modelos en software R, utilizando las variables con dependencia significativa. Con ello, se obtienen 18 diferentes modelos, la comparativa de estos se observa en la figura 1.

La medición de la exactitud y el riesgo, que se realiza convirtiendo los valores estimados en una variable dicotómica. Para ello, si el modelo refleja una probabilidad mayor a 0.5, se categoriza como

retención, por el contrario, se categorizó como una predicción de pérdida.

**Figura 1.**

*Comparativa de modelos de regresión logísticos*




*Nota.* Se aplica criterio de información de Akaike (AIC) y su exactitud de predicción. Elaboración propia, con Tableau.

Se observa cómo el modelo denominado "modelo 7", tiene el valor de exactitud más alto (69.39 % y se obtiene un valor para el criterio de información de Akaike (AIC) similar al de modelos (3,052.5). A continuación, su matriz de confusión.

**Tabla 2.**

*Matriz de confusión del modelo 7 de regresión logística*

| Valores estimados | Valores reales | | Total |
|---|---|---|---|
| | Perdido | Retenido | |
| Perdido | 1,058 | 463 | 1,521 |
| Retenido | 318 | 699 | 1,017 |
| Total | 1,378 | 1,162 | 2,538 |

*Nota.* Elaboración propia.

Dicho modelo está determinado por la siguiente combinación de variables en la notación informática utilizada por el software R.

***Estado del cliente - Segmento RFM***

***+ Método Envio***

***+ Cantidad Categorías*** **(1)**

***+ Sucursal + Término Pago***

***+ Segmento Cotización + RM***

## Discusión de resultados

Las variables RFM que se obtienen de la minería de datos demuestran tener correlaciones débiles, por lo que es factible la aplicación de algoritmos de agrupación. Con ello, se obtiene una segmentación de tres categorías, suficientemente heterogénea, con dependencia significativa ($p < 0.0001$) respecto a la variable de respuesta.

Además, se observa una relación débil entre las variables condición de pago ($p = 0.0276$) y canal de origen ($p = 0.4301$), con la variable de respuesta, respectivamente. Al analizar dicha falta de relación, se identifica que no hay suficientes observaciones para cada factor, haciendo que dichas variables no aporten información relevante al modelo.

El modelo de mejor ajuste presenta una exactitud de 69.39 %, sin embargo, no todos los errores de predicción representan riesgo para la empresa. Debido a las estrategias de servicio al cliente, únicamente los falsos positivos son riesgosos para la empresa. De dicho análisis, se obtiene que el riesgo de error de dicho modelo es de 12.53 %.

La medición de la exactitud del modelo se basa en la conversión de los valores estimados en una variable dicotómica asumiendo un umbral de 0.5. Al profundizar en el valor adecuado de dicha generalización, se obtienen exactitudes similares con valores de umbrales entre 0.45 y 0.55.

Jain et al. (2020) mencionan en los resultados de su estudio una exactitud de predicción de 85.23 % para su mejor modelo de regresión logística. Sin embargo, el estudio mencionado aplica este método en una empresa de telecomunicaciones, donde se tienen esquemas de contratos y otros datos de consumo que permiten tener más información del cliente.

De la misma forma, el estudio de Hargreaves (2019), que también se aplica al sector de telecomunicaciones, presenta una exactitud de predicción de 76.1 % aplicando una regresión logística binaria que incluye hasta 20 variables regresoras. A diferencia del autor, el presente estudio utiliza únicamente 7 variables regresoras para generar resultados similares (69.39 % de

exactitud), manteniendo el principio de parsimonia estadística.

## Conclusiones

1. Se segmenta a los clientes profesionales por sus variables RFM utilizando métodos de percentiles y de agrupación por K medias facilitando el análisis del comportamiento de los clientes.
2. Se identifica una correlación significativa entre la reciencia, monto promedio y antigüedad de cotización, mientras que las variables canal, frecuencia y condición de pago muestran una correlación débil.
3. En la construcción del modelo se obtiene una ecuación de regresión logística que permite estimar la retención o pérdida del cliente con un 69.39 % de exactitud, y un riesgo de 12.53 %.
4. Se construye el modelo de regresión logística que permite estimar la probabilidad de retención o pérdida de un cliente profesional, una base sólida para enfocar un sistema de seguimiento y servicio estratégico.

## Recomendaciones

A las empresas:

1. Implementar los grupos que genere el algoritmo de K medias para la segmentación de sus clientes.
2. Considerar la aplicación del modelo de regresión logística a otros segmentos de clientes para desarrollar estrategias de servicio.

## Referencias

Aleksandrova, Y. (2018). *Application of machine learning for churn prediction base don transactional data (RFM análisis)*. SGEM International Multidisciplinary ScientificGeoConference EXPO Proceedings. Congreso llevado a cabo en Sofia, Bulgaria. https://doi.org/10.5593/sgem2018/2.1/s07.016.

Cuadros, L. Gonzales, C. y Jiménez, P. (2017). Análisis multivariado para segmentación de clientes basada en RFM. *Tecnura, 21*(54), 41–51. https://doi.org/10.14483/22487638.12957

Dogan, O., Aycin, E., & Bulut, Z. (2018). Customer segmentation by using rfm model and clustering methods: a case study in retail industry. *International Journal of Contemporary Economics a n d Administrative Sciences, 8*(1), 1–19. https://doi.org/10.5930/ issn.1925-4423

Hargreaves, C. (2019). A Machine Learning Algorithm for Churn Reduction & Revenue Maximization: An Application in the Telecommunication Industry. *International Journal of Future Computer and Communication, 8*(4), 109–113. https://doi.org/10.18178/ijfcc.2019.8.4.550

Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn Prediction in Telecommunication using Logistic Regression and Logit Boost. *Procedia Computer Science,* 167, 101–112. https://doi.org/10.1016/j.procs.2020.03.187

## Información del autor

Ingeniero Mecánico Industrial, Pedro Pablo Morales Ortiz, Facultad de Ingeniería, Universidad de San Carlos de Guatemala, 2023.

Maestría en Artes en Estadística Aplicada, Universidad de San Carlos de Guatemala, 2023.

Afiliación laboral: Ferco.

Ingeniero Industrial, José Rolando Chávez Salazar, Facultad de Ingeniería, Universidad de San Carlos de Guatemala, 1996. Maestro en Administración Industrial, Universidad Rafael Landivar, 1999.

Afiliación laboral: Centro de Formación Profesional.