

Reto: Proceso estocásticos para modelar el comportamiento de los clientes en PiSA

Paulina Martínez López A01639743
Santiago Mora Cruz A01369517
Gabriel Reynoso Escamilla A01643561
Guillermo Villegas A01637169

Septiembre de 2024

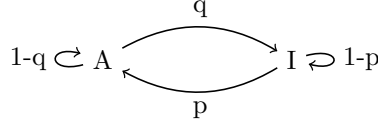
1 Introducción

Medicom es una empresa del grupo PiSA encargada de la distribución de productos a clientes de bajo volumen. Su misión es brindar el mejor servicio en distribución de medicamentos y otro equipo médico a través de su personal. La empresa busca un análisis de la información sobre la desactivación y reactivación de ventas de sus productos, es decir, si un producto se vende o no durante el plazo que se espera que sea comprado, así como el valor de vida de sus clientes. Para lograr esto se usará un modelo estocástico conocido como Cadena de Markov. Esta cadena se refiere a una secuencia de variables aleatorias, cuyos estados dependen únicamente del último estado en un tiempo pasado. [1] Para los fines de este artículo se aprovecharán algunas propiedades de dichas cadenas como lo son las probabilidades de transición entre estados, y la distribución estática de la cadena.

2 Metodología

Se comenzó con un Análisis Exploratorio de los datos (EDA), se exploraron las variables de la base de datos "fecha", "material", "id_cliente" y "venta". Para explorar las ventas se utilizó un diagrama basado en los cuartiles de los datos, diagrama de Cajas y Bigotes. Para cuantificar las ventas en periodo se realizaron histogramas de número de registros por año. Se identificaron valores negativos en la variable ventas, estos también se cuantificaron y se representaron con histogramas. Para conocer más a fondo este proceso consúltese el **Apéndice a.**

Para conocer explorar las probabilidades de reactivación y desactivación, tanto de materiales como de clientes, se propuso una cadena que podría usarse en ambos casos.



Esta cuenta con dos estados, Activo (A), e Inactivo (I). La probabilidad de desactivación es q , mientras que la de reactivación es p , esto lo podemos observar con la matriz de transiciones:

$$\mathcal{P} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

Por su estructura la cadena tiene características que nos pueden ayudar a dar respuesta a las preguntas del grupo PiSA, ambos estados están comunicados lo que hace su clase de comunicación la misma que su espacio de estados $\mathcal{E} = \mathcal{C} = \{A, I\}$, esto la hace irreducible. El periodo es el máximo común divisor del conjunto de número de pasos en el que es posible el estado regrese a si mismo, en el caso de la cadena propuesta:

$$d(\mathbf{A}) = d(\mathbf{I}) = \text{mcd}\{1, 2, 3, 4, \dots\} = 1$$

El periodo de ambos estados es uno esto hace de la cadena una cadena aperiódica. Por último se dice que es recurrente positiva porque si se inicia en cualquier estado, el tiempo esperado para volver al mismo es finito [1].

Estas características identifican a la cadena como ergódica, para la cual existe un vector de probabilidades estático, unico π , este vector es útiles porque nos permiten conocer es comportamiento a largo plazo tanto de clientes como materiales [2].

Para lograr un análisis más profundo se decidió analizar por separado los venta de los materiales y las compras de los clientes, sin embargo el proceso es análogo.

Para obtener la matriz de transición asociada a cada material o cliente, se extraen sus observaciones de la base de datos, filtrando el periodo de interés, enero 2021 hasta agosto de 2024. Después se agrupa esta información por mes y año sumando los valores de "ventas", en el caso de no existir registros en el periodo el valor es 0, este valor lo llamaremos X_t , siendo t el tiempo. Luego se definen los estados de la siguiente manera:

- Si $X_t > 0$: El estado es activo (A)
- Si $X_t = 0$: El estado es inactivo (I)
- Si $X_t < 0$:
 - Si $X_t < 0$ y $X_{t+1} > 0$: El estado es activo
 - Cualquier otro caso se considera inactivo

Para la matriz de transición P , se calcularon las probabilidades del siguiente modo: Se creó un data frame nuevo a partir de la clasificación de las transiciones mencionada anteriormente. En este data frame, con base en la fecha, se escribió para cada transición, cuantas veces ocurría. Con esto, se creó una matriz que tenía cuántas veces:

$$\begin{vmatrix} \text{Permaneció inactivo} & \text{Se reactivó} \\ \text{Se desactivó} & \text{Permaneció activo} \end{vmatrix}$$

Para calcular las probabilidades, se dividió cada valor entre la suma de su respectiva fila; obteniendo la matriz de transición.

Al ser ergódica podemos obtener la distribución estacionaria de la siguiente manera:

$$\pi = \left(\frac{p}{p+q}, \frac{q}{p+q} \right) \quad (1)$$

Para conocer la frecuencia de venta del producto podemos utilizar el tiempo medio de recurrencia, el número esperado de pasos para volver al estado de activación:

$$\mu_1 = \frac{1}{\pi_1} \quad (2)$$

Por último calculamos la matriz de transición al tiempo n .

Cabe mencionar que este proceso solo es útil para algunos materiales y clientes, existen dos casos en los que su probabilidad será distinta:

- Si el material o cliente, realizaron una sola transacción:

$$\mathcal{P} = \begin{pmatrix} 1 & 0 \\ N/A & N/A \end{pmatrix}$$

y,

$$\pi = (1, 0)$$

- Si el material o cliente, desde la primera transacción no se han desactivado:

$$\mathcal{P} = \begin{pmatrix} N/A & N/A \\ 0 & 1 \end{pmatrix}$$

y,

$$\pi = (0, 1)$$

También era de interés para grupo PiSA el valor de vida de los clientes. El *Customer Lifetime Value* (CLV) mide cuánto ingreso puede aportar considerando la duración de su actividad como cliente y el valor de sus transacciones. Por lo tanto, es una métrica que estima el valor total que un cliente generará para una empresa durante toda su relación con ella. Calculamos el CLV en un proceso escalonado donde primero debemos calcular el CLV monetario. Este nos ayuda a comprender el valor financiero básico que el cliente aporta a la empresa [3]. En un primer acercamiento lo vimos desde un punto de vista de marketing, se calculó utilizando la siguiente fórmula :

$$\mathbf{CLV}_{\text{monetario}} = \text{Promedio de ventas} \times \text{Número de transacciones}$$

$$\mathbf{CLV}_{\text{ajustado}} = \mathbf{CLV}_{\text{monetario}} \times \text{Tiempo de vida del cliente}$$

Donde el tiempo de vida del cliente son los días desde la primera hasta la última compra. Sin embargo decidimos tomar un acercamiento similar al de C.-J. Cheng y S.W. Chiu et. al. [4] y utilizar la probabilidad de desactivación p_{10} y valor promedio de ventas de cada cliente v y calcularla de la siguiente manera:

$$\text{CLV} = \mu_o v \quad (3)$$

Este valor permitirá conocer qué clientes aportan más a la empresa y por lo tanto es importante mantener buenas relaciones, y aquellos los cuales no aportan y sería conveniente incentivar ventas o cortar relaciones comerciales.

Con la finalidad de utilizar la información de la cadena y proporcionar una base para las estrategias de la empresa se diseñó un sistema de recomendación de productos usando la matriz de co-ocurrencia y la similitud del coseno, una métrica que cuando el valor se acerca al uno implica similaridad y al acercarse al cero disimilitud (en este caso no se han vendido juntos). Si tenemos una matriz de co-ocurrencia M , en la que cada entrada M_{ij} representa cuántas veces los productos i y j han sido comprados juntos, entonces cada fila (o columna) de esa matriz puede ser vista como un vector que describe las relaciones de un producto con los demás.

Así, el vector de cada producto es: P_i es $\vec{P}_i = (M_{i1}, M_{i2}, \dots, M_{in})$

Con estos vectores, se calcula la similitud del coseno entre dos productos P_i y P_j se define como:

$$\text{sim}(\vec{P}_i, \vec{P}_j) = \frac{\vec{P}_i \cdot \vec{P}_j}{\|\vec{P}_i\| \|\vec{P}_j\|}$$

Donde:

$\vec{P}_i \cdot \vec{P}_j = \sum_{k=1}^n M_{ik} M_{jk}$ es el producto punto.

$\|\vec{P}_i\| = \sqrt{\sum_{k=1}^n M_{ik}^2}$ es la norma del vector.

Con esto, la matriz de similitud S se construye calculando la similitud del coseno entre todos los productos:

$$S_{ij} = \frac{\sum_{k=1}^n M_{ik} M_{jk}}{\sqrt{\sum_{k=1}^n M_{ik}^2} \cdot \sqrt{\sum_{k=1}^n M_{jk}^2}}$$

Y para recomendar productos basados en uno dado P_i , encontramos los productos con mayor similitud en la fila i de la matriz S :

$\text{Recomendaciones}(P_i) = \text{Top } k \text{ productos } P_j \text{ tal que } S_{ij} \text{ es máximo, con } P_j \neq P_i$

Es posible observar la implementación de esta metodología en el **Apéndice b.**

3 Resultados

En el EDA se encontraron dos aspectos de los datos de importancia para el proyecto. El primero son los datos faltantes desde 2015 hasta 2020 (Fig. 1), supusimos que esto se debía a errores en el manejo de los datos, no a la falta de ventas, por lo que solo se realizó el modelo con la información disponible desde 2021. El otro son los valores negativos negativos en "ventas", se tuvo la opor-

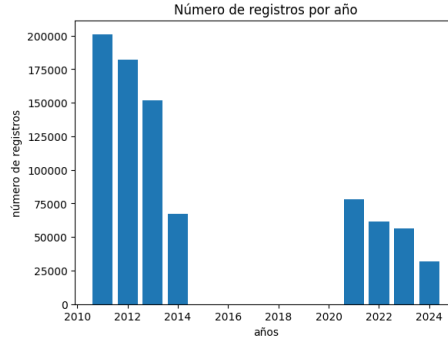


Figure 1: Número de registros por año

tunidad tener una sesión con los representantes de PiSA en la que se aclararon este punto, los valores negativos se refieren a devoluciones de materiales. Por último se supuso que la primera vez que se compró un material o un cliente compró coincidía con la primer fecha que se tenía en los datos de dicho cliente o material, a partir de 2021. La base de datos tiene registros diarios, sin embargo consideramos que sería de mayor utilidad agruparlos de manera mensual.

3.1 Materiales

Tomaremos como ejemplo del material 768:

- Última fecha en registro del material: 2024-08-06
- La matriz de transición asociada al material es:

$$\mathcal{P} = \begin{pmatrix} 0.9231 & 0.0769 \\ 0.0333 & 0.9666 \end{pmatrix}$$

- El vector de distribución estacionaria es: $\pi = (0.3023, 0.6977)$
- Dado que el material fue comprado, será comprado de nuevo en 1.43 meses

3.2 Clientes

Para el cliente 2222:

- Última fecha en registro para el cliente: 2024-02-29

- La matriz de transición asociada al cliente es:

$$\mathcal{P} = \begin{pmatrix} 0.7059 & 0.2941 \\ 0.2308 & 0.7692 \end{pmatrix}$$

- El vector de distribución estacionaria es: $\pi = (0.4397, 0.5603)$
- Dado que el cliente compró, comprará de nuevo en 1.78 meses

4 Análisis

El análisis realizado tanto con la cadena de Markov como con el CLV permitirán a Medicom: Conocer los productos con ventas más lentas y periodos de inactividad más largos, lo que será útil para idear una estrategia de mayor movilización del producto o retirarlo del mercado. También dará información del comportamiento de los clientes permitiendo confiar en los que tienen un estado estable de activación, sin descuidarlos, y prestar mayor atención en aquellos que tengan un tiempo entre compras más largos.

Para los clientes, teniendo los tiempos medios de recurrencia, el CLV y las probabilidades, se puede hacer una ponderación de estas características para cada cliente y, basado en esto, ofrecer beneficios a los clientes frecuentes y a aquellos que son recurrentes para lograr dos cosas: evitar que pasen a un estado inactivo e incentivar a nuevos clientes a comprar más frecuentemente.

Como estrategia proponemos lo siguiente: Para **Materiales**, segmentarlos en tres casos: Si el material se ha comprado cada mes desde su primer registro, se recomienda mantener el inventario de este bien abastecido, y promocionarlo como *best-seller*. Si el producto tiene ventas intermitentes se recomienda agruparlo en promociones de venta cruzada con productos que se vendan de manera similar, además de ajustar el inventario a su demanda variable. Por último si el producto tiene ventas irregulares, se recomienda ofrecerlo a clientes que lo han comprado antes con ofertas especiales.

Los **Cientes** también se separarán en tres casos: Si el cliente es leal (tiene compras mensuales desde su primera activación), se recomienda ofrecer un programa de fidelidad. Si el cliente tiene compras intermitentes se recomienda implementar recordatorios automáticos ajustados a su tiempo medio de recurrencia. Si las compras del cliente son más espaciadas se recomienda mantener contacto con el cliente, mediante promociones, comunicados, etc., para intentar que compre más frecuentemente.

5 Conclusiones

Con este trabajo fue posible crear una interfaz en la que la empresa, ingresa el material o cliente de su interés y regresa la tasa de desactivación y reactivación, así como sus probabilidades, también el tiempo medio entre compras. Además

en el caso que sea un cliente regresa su CLV. Junto con esto proporciona las posibles estrategias comentadas anteriormente. Aunque la aplicación creada es útil es posible utilizar el modelo para clasificación de clientes, predicción de demanda o simplemente otra clase de estrategias de la mano de un equipo de marketing o ventas. Además se podría agregar información para mejorar o aumentar el modelo como los datos faltantes, el número de unidades vendidas, o las categorías de los productos.

References

- [1] W. K. Ching and M. K. Ng, *Markov chains : models, algorithms and applications*. Springer, 2006.
- [2] R. G. Gallager, *Discrete Stochastic Processes*. New York, NY: Springer, 1996.
- [3] Qualtrics, “Customer lifetime value (clv): qué es y cómo calcularlo.” <https://www.qualtrics.com/es-es/gestion-de-la-experiencia/cliente/customer-lifetime-value/>. Accessed: 2024-08-24.
- [4] C.-J. Cheng, S. Chiu, C.-B. Cheng, and J.-Y. Wu, “Customer lifetime value prediction by a markov chain based data mining model: Application to an auto repair and maintenance company in taiwan,” *Scientia Iranica*, vol. 19, no. 3, pp. 849–855, 2012.