Predicción de la rotación de clientes de TDC mediante un Bosque de Supervivencia Aleatorio

Ángel Gustavo José Martínez

1 de febrero de 2024

Contexto

La rotación de clientes con TDC se refiere a cuando los clientes cancelan o dejan de usar su tarjeta de crédito, lo cual conduce a la pérdida de ingresos. Implementar una estrategia proactiva para identificar los factores y clientes con altas probabilidades de rotación, brindan muchas posibilidades para retenerlos.

Incorporar el Análisis de supervivencia es de gran utilidad para la gestión de la retención debido a que considera la censura, en particular, censura por la derecha. Es decir, considera que solo por que el evento (rotación) no ha ocurrido en el periodo elegido, no garantiza que no se producirá. En combinación con un clasificador como el bosque aleatorio, nos permite hacer predicciones individuales con base en las características de los clientes.

Se utiliza el conjunto de datos llamado *Bankchurners* el cual contiene información sociodemográfica y de transacciones de clientes con TDC de un Banco. Lo puedes encontrar en el siguiente link:
https://leaps.analyttica.com
La descripción de las columnas se encuentra en la parte final de este documento.

Algoritmo

El algoritmo general proporcionado por Ishwaran et al, es el siguiente:

1. Extraer B muestras del mismo tamaño con remplazo del conjunto de entrenamiento.
2. Para cada muestra $b = 1, ..., B$ construir un árbol de supervivencia.
   - *a*) En cada nodo, seleccionar un subconjunto aleatorio de predictores tal que proporcione los nodos hijos que maximicen la diferencia en la función objetivo.
   - *b*) Repetir a) recursivamente en cada nodo hijo hasta que se cumpla un criterio de parada.
3. Calcular la función acumulativa de riesgo para los B árboles y promediarlas para obtener el ensamble de la función acumulativa de riesgo.

Resultados obtenidos

C-Index

Concordance Index o *C-Index*, es una generalización del área bajo la curva ROC que considera tiempos de supervivencia, por lo cual, se utiliza para la evaluación del desempeño del modelo.

$$C - index = \frac{\sum_{i,j} 1_{T_J<T_i} \cdot 1_{\eta_j>\eta_i} \cdot \delta_j}{\sum_{i,j} 1_{T_J<T_i} \cdot \delta_j},$$

donde:

- η_i es el score de riesgo de rotación predicho de la unidad i.
- $1_{T_J<T_i} = 1$ si $T_J < T_i$, 0 en otro caso.
- $1_{\eta_j>\eta_i} = 1$ si $\eta_j, > \eta_i$, 0 en otro caso.

Similar a AUC, $C-index = 1$ corresponde a una predicción perfecta y un $C-index = 0{,}5$ corresponde a una predicción aleatoria.

En este caso:

- $C - index = 0{,}96$ en datos de entrenamiento.
- $C - index = 0{,}93$ en datos de validación.

Lo cual indica un modelo con excelente desempeño.

Predicciones

La predicción del score de rotación con base en las características del cliente **representa el promedio del número esperado de eventos sobre todos los nodos terminales de cada uno de los árboles de supervivencia.** Para calcularla se usan los estimadores no paramétricos de Kaplan-Meier y Nelson-Aalen.

Se toman 5 muestras del conjunto de validación para predecir su score de riesgo de rotación, su función de supervivencia y función de riesgo acumulado:

muestra1: **11.35**; muestra2: **3.8**; muestra3: **0.88** ; muestra4: 0.83; muestra5: 35.44.

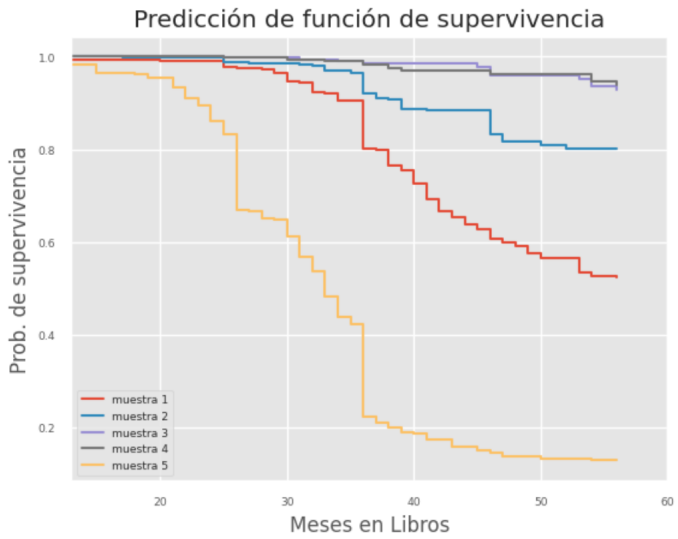


Figura 1: *Predicción de la función de sup. para 5 muestras en conjunto de validación.*

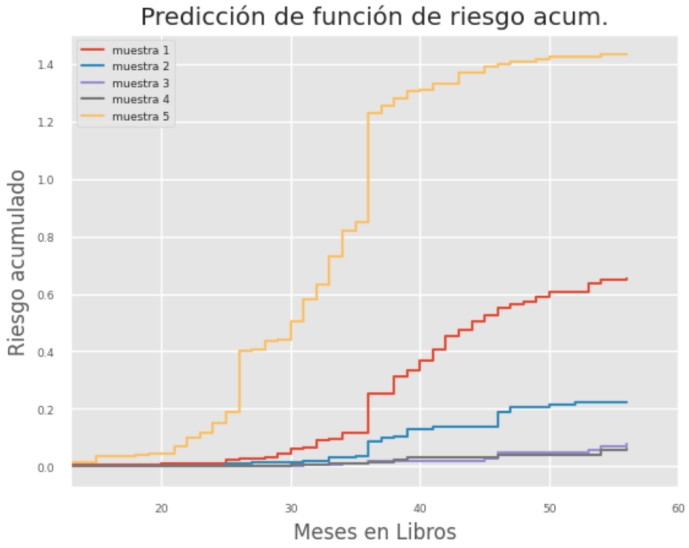


Figura 2: *Predicción de la función acumulada de riesgo para 5 muestras en conjunto de validación.*

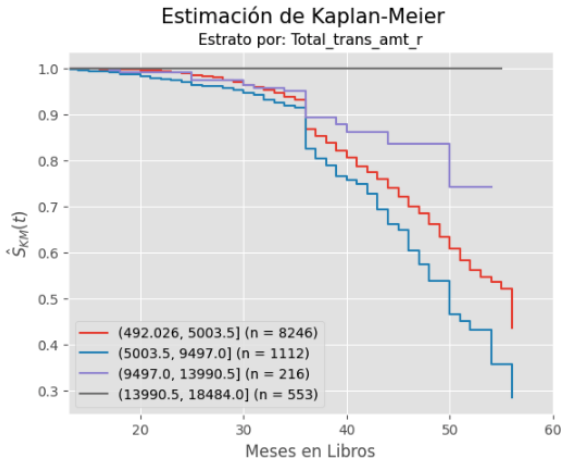
La muestra5 tiene el score más alto, esto también se ve relejado en su función de supervivencia y riesgo. Tiene sentido, pues **este cliente rotó a los 32 meses en libros.**

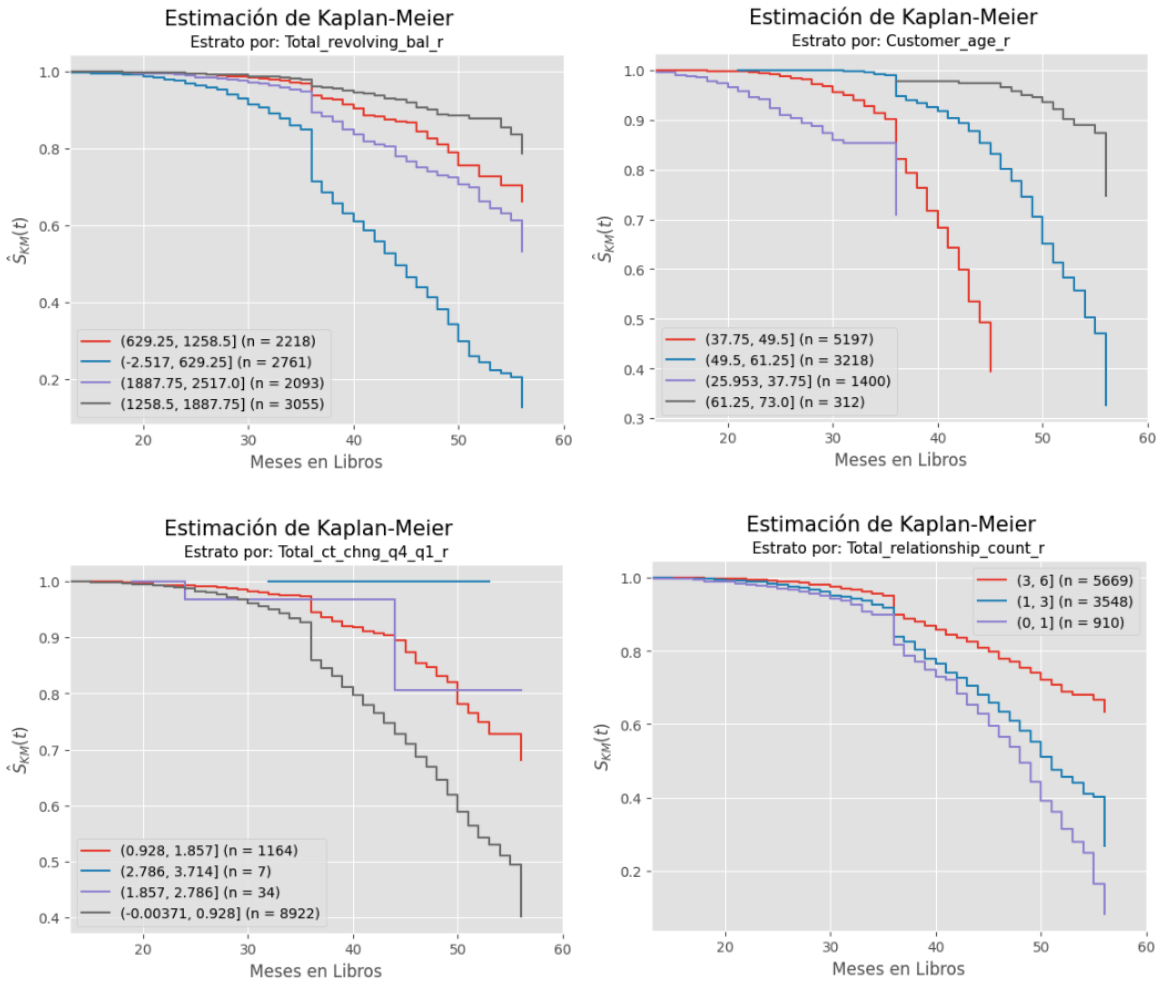
Los 5 factores más importantes que afectan a la rotación

De acuerdo al modelo ajustado y a *Permutation feature importance*, son:

1. **Total Trans Amt**: Saldo de transacciones en los últimos 12 meses.
2. **Total Revolving Bal**: Saldo revolvente.
3. **Customer Age**: Edad del cliente.
4. **Total Ct Chng Q4 Q1**: Relación entre el recuento total de transacciones en el cuarto trimestre y el recuento total de transacciones en el primer trimestre.
5. **Total Relationship Count**: Número de productos del cliente.

Graficando la función de supervivencia para estos factores y divididos por estratos, se tiene:





De acuerdo a estas estimaciones, un segmento evidente, son aquellos clientes que:

- Los saldos de sus transacciones durante 12 meses están entre 5k y 10k.
- Tienen saldos revolventes menores a 630.
- Clientes entre 26 y 38 años.
- Tienen solo un producto en el Banco.

Sin embargo, para crear un segmento prioritario robusto, debe construirse con el score de riesgo de rotación, pues este último captura la interacción entre todas las características consideradas.

Contacto: angujoma@gmail.com

Link al código utilizado: GitHub

Anexo: Información usada

Descripción de las columnas:

- CLIENTNUM: Identificador del cliente
- Attrition Flag: Indica si la cuenta del cliente está cerrada o activa.
- Customer Age: Edad en años.
- Gender: Género.
- Dependent count: Número de dependientes.
- Education Level: Nivel escolar.
- Marital Status: Estado civil.
- Income Category: Ingreso por categoría.
- Card Category: Categoría de la TDC.
- Months on book: Meses en libros.
- Total Relationship Count: Número de productos del cliente.
- Months Inactive 12 mon: Meses inactivo en los últimos 12 meses.
- Contacts Count 12 mon: Número de contactos entre el banco y el cliente en los últimos 12 meses.
- Credit Limit: Límite del crédito
- Total Revolving Bal:Saldo revolvente.
- Avg Open To Buy: Promedio de 12 meses de saldo por usar de la TDC.
- Total Trans Amt: Saldo de transacciones en los últimos 12 meses.
- Total Trans Ct: Número de transacciones en los últimos 12 meses
- Total Ct Chng Q4 Q1: Relación entre el recuento total de transacciones en el cuarto trimestre y el recuento total de transacciones en el primer trimestre
- Total Amt Chng Q4 Q1: Relación entre el monto total de la transacción en el cuarto trimestre y el monto total de la transacción en el primer trimestre
- Avg Utilization Ratio: Representa cuánto del crédito disponible gastó el cliente.

Referencias

Ishwaran H, Kogalur U, Blackstone E, Lauer M. Random survival forests. The Annals of Applied Statistics. 2008; 2(3):841–860.

Wright, Marvin N., Theresa Dankowski and Andreas Ziegler. "Random forests for survival analysis using maximally selected rank statistics."" Statistics in medicine 36 8 (2017): 1272-1284.

PySurvival User Guide.

sksurv User Guide.