Análisis de clasificación para una red de corresponsales bancarios en Perú utilizando series de tiempo y datos categóricos

Trabajo de grado para la Maestría en Inteligencia Analítica para la Toma de Decisiones
Universidad de Los Andes - Departamento de Ingeniería Industrial

Estudiante: Edgar Andrés García Hernández - 200512532
Asesor: Felipe Montes Jiménez
Preasesora: Karen Daniela Angulo Díaz
30 de noviembre de 2018

Resumen

Una red peruana de corresponsales bancarios, los cuales son comercios habilitados para prestar ciertos servicios financieros a nombre de entidades financieras, busca mejorar su rentabilidad y analizar mejor el comportamiento de sus puntos a lo largo del tiempo, para lo cual solicitó un análisis acerca de la evolución de las transacciones en su red y la relación entre dichos comportamientos transaccionales con el estado actual de sus puntos. Para ello, se empleó un análisis de clúster para series de tiempo sobre datos transaccionales para una red de corresponsales bancarios en Perú a lo largo de tres años, complementado con un *random forest* para asociar dichos clústeres con características asociadas a los corresponsales, y un modelo de máquina de vectores de soporte (SVM) para relacionar dichos clústeres con el estado actual de los puntos. Se hallaron tres patrones diferentes de comportamiento transaccional, los cuales fueron relacionados exitosamente con dichas características y el estado del corresponsal. A partir de lo anterior, se observó que la red de corresponsales bancarios aún no ha logrado recuperar los niveles transaccionales de hace tres años, al tiempo que necesita expandir sus criterios de selección para incluir tipos de comercio y tener más en cuenta la actividad principal del mismo. Igualmente, se apreció cómo el sector donde está ubicado el comercio afecta el resultado del mismo. Lo anterior agrega valor por medio de insumos cuantitativos para la toma de decisiones de apertura de puntos futuros.

Problema de negocio

Introducción

Los corresponsales bancarios se han constituido a lo largo del tiempo en una forma de ofrecer servicios bancarios a zonas rurales o de bajos ingresos, cuyos altos costos de operación y bajos volúmenes transaccionales las convierten en poco atractivas para la llegada de oficinas bancarias tradicionales (Jayo, 2010). Dicho canal ha demostrado su utilidad para la bancarización en países como Brasil, Perú y Colombia, donde se han logrado altos grados de cobertura bancaria, facilitando el desembolso de programas sociales gobierno a persona (G2P) (Loureiro, Madeira, & Bader, 2011) y aumentando el número de personas con un producto bancario.

Sin embargo, la sostenibilidad de este canal no siempre está garantizada. En Colombia, el canal ha requerido de subsidios para asegurar la llegada de este a todos los municipios del país, creando dudas sobre su sostenibilidad a largo plazo (Cardona Prada, 2017). Por otro lado, en Brasil se ha observado cómo los corresponsales han servido para sustituir oficinas bancarias en ciertos segmentos (Loureiro etal., 2011). Relacionado con lo anterior se puede hallar la rentabilidad del canal, la cual depende directamente del número de transacciones en los puntos de corresponsalía, los productos ofrecidos y la demanda por parte de los clientes.

A partir de lo anterior, en el presente estudio se realizó una aplicación del análisis de clúster de series de tiempo transaccionales como técnica para analizar el comportamiento de los corresponsales bancarios afiliados a la red de RedCB, buscando identificar patrones a lo largo de su funcionamiento. Igualmente, se relacionaron dichos grupos con características descriptivas de los corresponsales bancarios con la finalidad de identificar mejor estos grupos a partir de las mismas utilizando un modelo *random forest*, y finalmente la relación entre el clúster de clasificación y el estado actual del corresponsal por medio de un modelo de máquina de vectores de soporte (SVM). Con lo anterior se busca identificar mejor el comportamiento de

los corresponsales, las diferencias entre corresponsales exitosos y no exitosos, y convertirse en un insumo para la selección futura de puntos.

Caso de estudio: RedCB

RedCB es una compañía que ofrece servicios de administración y agregación de redes de corresponsalía bancaria, una modalidad contractual bajo la cual es posible ofrecer servicios bancarios en comercios tradicionales. Bajo dicho sistema, el administrador de la red (RedCB) actúa como un intermediario entre los bancos y los comercios, haciéndose cargo de los acuerdos contractuales con el banco para definir los productos a ser ofrecidos por el canal, seguido por el reclutamiento, provisión tecnológica, operación y soporte técnico de los servicios de corresponsalía bancaria ofrecidos por los comercios.

La compañía trabaja con el banco Banco1 y la financiera FinBanco1, con una red de 450 corresponsales bancarios activos en agosto de 2018, distribuidos en múltiples ciudades del Perú. La mayoría de los puntos se localiza en Lima Metropolitana, compuesta por las provincias de Lima y Callao, siguiendo la tendencia de concentración económica y poblacional en dicha región del país. En la actualidad, dichos corresponsales bancarios ofrecen servicios agrupados en cuatro grandes categorías: pagos de facturas de terceros y obligaciones bancarias, recargas de telefonía móvil y tarjetas de transporte público, depósitos a cuentas bancarias, retiros de cuentas bancarias y consultas a cuentas bancarias o tarjetas de transporte público. En promedio, cada corresponsal ejecuta aproximadamente 561 transacciones por mes.

Dicho volumen de transaccionalidad representa un promedio bajo, explicado parcialmente por una decisión estratégica de RedCB en 2013, bajo la cual se hizo énfasis en la expansión de la red de corresponsales bancarios, en detrimento del número de transacciones por punto. Lo anterior afecta los márgenes obtenidos por la compañía, debido a los costos fijos incurridos al montar cada punto de corresponsalía bancaria, tales como equipos, conexión a Internet y entrenamiento de personal en los comercios aliados.

Dicha decisión estratégica fue modificada en 2015, cuando se comenzó a priorizar el número de transacciones por punto. Lo anterior, si bien llevó a un aumento marcado en el número de transacciones por corresponsal, implicó un alto número de cierres en la red. Y si bien dicho aumento fue significativo, aún no se han cumplido las expectativas de la compañía, la cual apunta a ejecutar un mínimo de 1200 transacciones por punto. La remuneración a dichos puntos corre por cuenta de una estructura contractual definida entre RedCB y el banco, bajo la cual tanto el corresponsal como RedCB obtienen una comisión por transacción. Dichas comisiones son fijas, independientemente del monto de la misma. Igualmente, RedCB ha estado buscando concretar acuerdos para ofrecer servicios de corresponsalía con otros bancos peruanos, cuyos nombres se mantienen ocultos por razones de confidencialidad empresarial.

Desde hace unos años, RedCB recibió una inyección de capital de CapitalCO, una compañía colombiana de la industria gráfica, para la operación y despliegue de dicha red de cajeros corresponsales. Sin embargo, las utilidades por dicha inversión aún no son las esperadas. Al mismo tiempo, CapitalCO contrató un equipo de consultoría para servir de puente y supervisor de las operaciones de corresponsalía bancaria de RedCB, dada su falta de experiencia.

Ante lo anterior, el equipo de consultoría buscó mejorar la rentabilidad de los puntos de RedCB. Para ello, ha decidido enfocarse en los servicios prestados en cada corresponsal, así como en el análisis de patrones en los comportamientos de los corresponsales. Con lo anterior, el equipo busca obtener proyecciones adecuadas para el comportamiento de cada punto, así como identificar comportamientos comunes a corresponsales que están a punto de abandonar la red de corresponsalía. El análisis de datos hizo parte integral de un plan diseñado por el equipo de consultoría para mejorar la operación y rentabilidad de los corresponsales bancarios de RedCB.

Preguntas de negocio

Para el presente estudio, y dadas las necesidades de RedCB, las preguntas de negocio elegidas fueron las siguientes:

1. ¿Cuáles son las categorías o grupos de corresponsales bancarios en la red de RedCB, según sus volúmenes transaccionales y su comportamiento en el tiempo?
2. ¿Cuáles son las características más relevantes para diferenciar los corresponsales bancarios activos en la red de RedCB de los corresponsales bancarios deshabilitados en dicha red?
3. ¿Cuáles serían las recomendaciones para las características de los comercios a seleccionar para

la red, buscando maximizar el volumen transaccional de los corresponsales bancarios?

Objetivos del proyecto

Con este proyecto se buscan identificar patrones en el tiempo en los volúmenes transaccionales de los corresponsales bancarios, buscando clasificar estos corresponsales en grupos por medio de métodos de clasificación no supervisada. A partir de lo anterior, se buscarán relacionar estos clústeres con la probabilidad de cierre de un punto, utilizando un método de clasificación supervisada por SVM. Finalmente, se analizarán diferencias entre los clústeres transaccionales para otras variables, como su ubicación y su actividad principal.

A partir de lo anterior se busca obtener una mejor comprensión de la red de corresponsalía bancaria de RedCB, y los criterios que determinan el éxito de estos corresponsales. También se busca analizar el comportamiento de los puntos a lo largo del tiempo, para los productos ofrecidos. Buscando ofrecer una respuesta a las necesidades del negocio, se entregaron al equipo de consultoría y a la gerencia de RedCB el presente informe, así como los resultados del mismo. Lo anterior se convertirá en un insumo para futuras tomas de decisiones en cuanto a los criterios de selección para la apertura de nuevos puntos de corresponsalía bancaria.

Actores involucrados

El presente proyecto será de interás para la toma de decisiones y operaciones en la red de corresponsalía bancaria de RedCB, la cual está asociada a las siguientes partes interesadas:

CapitalCO

Compañía de la industria gráfica colombiana, la cual ha invertido capital propio en RedCB con la finalidad de diversificar sus líneas de negocio. Busca maximizar la rentabilidad de su inversión en la red de corresponsalía bancaria de RedCB.

RedCB

Operador de la red de corresponsalía bancaria analizada, a cargo de actividades de reclutamiento de corresponsales, enlaces con los bancos, definición de productos ofrecidos en la red y seguimiento de operaciones. Busca optimizar sus operaciones e identificar patrones que le permitan mejorar su proceso de toma de decisiones al momento de la apertura de nuevos corresponsales bancarios.

Asesores especializados

Un equipo de consultoría contratado por CapitalCO con la finalidad de asesorar en la operación y la toma estratégica de decisiones en la red de corresponsalía bancaria de RedCB. Busca ofrecer soluciones para mejorar la operación y rentabilidad de los corresponsales bancarios de la red, y así obtener una mayor rentabilidad para CapitalCO

Bancos

Además de las tres partes ya mencionadas que se encuentran involucradas directamente con la operación de la red, se encuentran los bancos participantes en la red de corresponsalía. Por un lado se encuentra el banco Banco1, el cual ofrece sus servicios actualmente en la red. Una mejor selección de corresponsales producto de mejores decisiones por parte de RedCB a partir de esta información podría ayudar en sus objetivos de mejorar su posición competitiva y obtener un mayor alcance geográfico para sus operaciones a nivel nacional.

Por otro lado, se encuentran otros bancos peruanos que actualmente no ofrecen sus servicios por medio de la red de RedCB, pero podrían estar interesados en acceder a la misma en el futuro. Nuevamente, se observa como la posibilidad de mejores decisiones en cuanto a apertura de corresponsales y mejores resultados transaccionales podrían resultar atractivos para aumentar su alcance geográfico a un menor costo, aumentando su contacto con clientes existentes y potenciales.

Limitaciones y delimitaciones

Para este proyecto, se tuvieron en cuenta únicamente las transacciones ejecutadas a partir de junio de 2015. Esto se debe principalmente al cambio en la estrategia comercial por parte de RedCB, bajo el cual se enfatizó en el número de transacciones por corresponsal bancario sobre el tamaño de la red.

Igualmente se analizaron únicamente puntos en Lima Metropolitana, comprendiendo las provincias de Lima y Callao. Esta región concentra más del 80 % de los puntos de RedCB a nivel nacional, si bien la red tiene presencia fuerte en ciudades intermedias como Trujillo.

Finalmente, no fue posible obtener una base de datos con puntos identificados por RedCB como candidatos a convertirse en corresponsales bancarios. Lo anterior precluyó la posibilidad de un ejercicio de tipo prescriptivo, el cual se propone como investigación futura.

Metodología

En la siguiente sección, se mostrarán cuáles son los datos y variables disponibles para resolver el problema, así como las técnicas a estadísticas a utilizar y las herramientas de software.

Información disponible

Para el presente estudio, se emplearon dos conjuntos de información suministrados por RedCB. El primer conjunto consiste en una base de datos con los registros para cada una de las transacciones efectuadas en la red de corresponsales bancarios, entre el 1 de junio de 2015 y el 31 de agosto de 2018. Los campos en dicho conjunto son descritos en el Cuadro 1.

Cabe mencionar que, en cuanto a tipos de transacción, existían originalmente 27 tipos de transacción. Dichos tipos fueron agrupados en cuatro categorías según el criterio del equipo de consultoría, resultando en una clasificación mostrada en el Cuadro 2.

Para los propósitos del presente estudio, el código de los comercios fue elegido como la variable de identificación. Los datos suministrados por RedCB fueron transformados en R, utilizando los paquetes *reshape2* (Wickham, 2007), *dplyr* (Wickham, François, Henry, & Müller, 2018) y *xts* (Ryan & Ulrich, 2018). Dichos paquetes fueron empleados para transformar la información original a series de tiempo, bajo la cual cada fila representaba un corresponsal bancario a partir de su código interno, y cada columna representaba una fecha específica continua, del 1 de junio de 2015 al 31 de agosto de 2018. En caso que no hubiesen registros de transacciones de un tipo para un comercio en un día, dichos valores faltantes fueron reemplazados por ceros.

Finalmente, se utilizó como variable medida dentro de las series de tiempo el número de transacciones ejecutado por un corresponsal bancario en un día. Con el fin de obtener una mejor diferenciación, se crearon 2319 matrices separadas, una por cada una de los códigos únicos existentes en la base de datos. Para cada una de dichas matrices, las fechas conformaban los nombres de las filas, mientras las columnas representaban los cuatro tipos de transacciones mostrados en el Cuadro 2.

A su vez, el segundo conjunto de datos suministrado por RedCB contiene los comercios que ejercen o alguna vez ejercieron como corresponsales bancarios en la red. Los campos en dicho conjunto de datos son descritos en el Cuadro 3.

Técnicas estadísticas a utilizar

Con el fin de resolver las preguntas de negocio, una de las labores más importantes consiste en agrupar los corresponsales bancarios pertenecientes a la red, buscando obtener patrones y características comunes que permitan mejorar el entendimiento de la red de corresponsales a lo largo del tiempo. Para ello, se propone el uso de una técnica de análisis de clúster sobre series de tiempo.

El análisis de clúster es una técnica que permite agrupar observaciones en grupos u objetos homogéneos a partir de características comunes, viendo aplicaciones en ciencias tanto naturales como sociales (Hair etal., 2006). Se ha utilizado tanto en estudios sobre banca (Sørensen & Puigvert Gutiérrez, 2006) y comercios al por menor (Ramakrishnan, 2010), los dos campos más relacionados con la corresponsalía bancaria como canal. Dada esta versatilidad, esta técnica se constituye en una herramienta útil para identificar grupos y tipos de corresponsal en la red, destacando su naturaleza como técnica de clasificación no supervisada.

Sin embargo, los datos utilizados en el presente estudio seguían un formato de series de tiempo. Según Roelofsen (2018), las técnicas de análisis de clúster tradicionales como k-medias no deben utilizarse para series de tiempo, debido al uso de las distancias euclidianas como criterio de agrupación de los datos. Dichas distancias no son un criterio de medición adecuado para series de tiempo, en las cuales los valores no son independientes entre sí.

Debido a lo anterior, se utilizaron técnicas de análisis de clúster para series de tiempo, con distancias especiales. Según Liao (2005), existen tres tipos de técnicas de análisis de clúster para series de tiempo: (a) agrupación basada en la forma, obtenida a partir de los valores originales de los datos, sin ningún tipo de transformación; (b) agrupación basada en características, a partir de vectores de características obtenidos

Cuadro 1: Listado de variables en la base de datos transaccional

Variable	Descripción	Tipo
Código	Código único del corresponsal bancario que ejecutó la transacción	Numérico
Punto de venta	Nombre del comercio que opera como corresponsal bancario en la red	Caracter
Fecha de operación	Tipo de operación ejecutada. Puede ser de cuatro tipos	Fecha
Tipo de operación	Monto transado en la operación	Caracter
Monto	Hora de ejecución de la transacción	Numérico
Hora	Distrito donde se encuentra ubicado el punto de venta	Hora
Distrito	Provincia donde se encuentra ubicada el punto de venta	Caracter
Provincia	Departamento donde se encuentra ubicado el punto de venta	Caracter
Departamento	Código único del corresponsal bancario que ejecutó la transacción	Caracter

Cuadro 2: Categorías transaccionales y miembros de las mismas

Transacción	Recargas	Recaudos	Servicios de información	Financieras
Recargas de Bitel	X			
Recargas de Claro	X			
Recargas de Entel	X			
Recargas de Movistar	X			
Pago de servicio Claro		X		
Pago de servicio Sedapal		X		
Pago de servicio Telefúnica Básica		X		
Pago de servicio Telefúnica Cable Mágico		X		
Pago de servicio Telefúnica Celular		X		
Pago de convenios institucionales		X		
Movimientos cuenta de ahorros			X	
Movimientos cuenta corriente			X	
Movimientos tarjeta de crédito			X	
Saldo tarjeta de crédito			X	
Saldo tarjeta de transporte integrado			X	
Consulta préstamos personales			X	
Depósito cuenta de ahorro propias				X
Depósito cuenta de ahorro de terceros				X
Depósito cuenta corriente propia				X
Depósito cuenta corriente de terceros				X
Pago tarjeta de crédito				X
Pago préstamo personal				X
Retiro cuenta de ahorros				X
Retiro cuenta corriente				X

Cuadro 3: Listado de variables en la base de datos de puntos históricos

Variable	Descripción	Tipo
Código	Código único del corresponsal bancario	Numérico
Dirección	Dirección donde se encuentra ubicado el corresponsal bancario	Caracter
Distrito	Distrito donde se encuentra ubicado el punto de venta	Caracter
Provincia	Provincia donde se encuentra ubicada el punto de venta	Caracter
Departamento	Departamento donde se encuentra ubicado el punto de venta	Caracter
Estado	Estado del corresponsal, puede estar activo o deshabilitado	Binario

de los datos originales; y (c) agrupación basada en modelos, a partir de parámetros de modelos ajustados sobre los datos originales.

Para el presente trabajo, se utilizó un clúster particional basado en la forma de las series de tiempo analizadas, utilizando una medida de distancia denominada *dynamic time warping* (DTW). Según Sardá-Espinosa (2017), dicho algoritmo compara dos series de tiempo y busca obtener el camino óptimo de alineación entre ellas, sin requerir que los patrones coincidan perfectamente en el tiempo. La Figura 1 demuestra visualmente dicha propiedad.

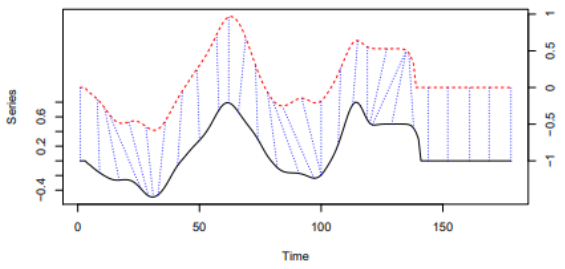


Figura 1: Alineación de series de tiempo obtenida por el algoritmo DTW. Fuente: Sardá-Espinosa (2017)

Sin embargo, este camino de alineación está sujeto a las siguientes restricciones (E. J. Keogh & Pazzani, 2000):

- Condiciones de frontera: el camino de alineación debe comenzar y finalizar en celdas opuestas de la matriz. En otras palabras, las fechas de inicio y de fin de las series comparadas deben coincidir.
- Continuidad: las series de tiempo comparadas deben ser continuas, no pueden presentar interrupciones.
- Monotonicidad: únicamente se permite un valor para un momento de tiempo en cada una de las series.

Utilizando dicho algoritmo para calcular las distancias entre observaciones, se calcularon un análisis de clúster jerárquico y un análisis de clúster particional para los datos transaccionales obtenidos. Una de las principales ventajas de dicho algoritmo es la posibilidad de trabajar con series de tiempo multivariadas, lo cual sirvió para los propósitos del presente estudio. Sin embargo, esta medida de distancia presenta un gran problema: altas exigencias computacionales, debido a que sus requisitos de tiempo van creciendo de forma cuadrática junto con el número de observaciones y variables analizadas (Roelofsen, 2018).

Buscando aumentar la eficiencia computacional tanto para el presente estudio como para aplicaciones futuras, fue necesario reducir las dimensiones. Lo anterior se realizó utilizando una técnica de reducción de dimensiones denominada *piecewise aggregate approximation* (PAA), propuesta por Keogh, Chakrabarti, Pazzani, y Mehrotra (2001). Bajo dicha técnica, la serie de tiempo es dividida en un número predeterminado de "cuadros" de igual tamaño, dentro de los cuales se calcula el valor promedio de la variable observada:

$$\bar{X}_i = \frac{N}{n} \sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i} X_j$$

En la ecuación anterior, $\bar{X}_i$ indica el valor i dentro de una serie de tiempo $\bar{X}$ reducida obtenida a partir de esta técnica de clasificación, la cual tiene una longitud igual al número de "cuadros" N elegido, el cual tendrá un tamaño j. Este valor se obtiene a partir de una serie de datos con longitud n, promediando dichas observaciones. Un ejemplo visual de una serie de tiempo original y una serie transformada se puede observar en la Figura 2.

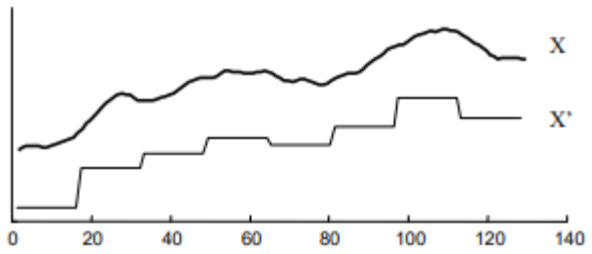


Figura 2: Ejemplo de aplicación de la técnica de reducción de dimensiones PAA. Fuente: E. Keogh et al. (2001)

Una vez obtenida esta agrupación para los corresponsales bancarios de la red, se ejecutó un modelo de máquina de vectores de soporte (SVM) con la finalidad de clasificar el estado actual del corresponsal a partir del clúster transaccional al cual este pertenece. Dicho modelo fue elegido debido a su capacidad de ejecución y velocidad, además de la posibilidad de trabajar con múltiples formas funcionales para el *kernel* generador de observaciones. Igualmente, el tamaño de la muestra y el bajo número de variables llevó a que fuese innecesario el utilizar redes neuronales.

Finalmente, se empleó un modelo de clasificación *random forest* para relacionar el distrito y la actividad principal del punto con su respectivo clúster transaccional. Los motivos para elegir dicho modelo consistieron en su correcto desempeño con múltiples variables categóricas, así como su capacidad de evitar sesgos excesivos en caso de variables correlacionadas (James, Witten, Hastie, & Tibshirani, 2013). Igualmente, el tamaño de la muestra no asistía en la ejecución de un modelo de redes neuronales, el cual no obtuvo un mejor desempeño que este modelo bajo la métrica elegida (AUC). Finalmente, destacó la posibilidad de verificar la importancia de las variables en este modelo, dándole un componente de interpretabilidad no disponible en un modelo de redes neuronales, al tiempo que garantiza un mayor grado de precisión frente a un modelo de regresión logística.

Como paso previo a estos dos modelos, se extrajo la actividad del comercio por medio de minería de texto y extracción de raíces principales sobre los nombres de los mismos. Igualmente, debe mencionarse que debido al desbalance en la muestra para las categorías de clústeres halladas, fue necesario utilizar remuestreo tipo *Synthetic Minority Over-sampling Technique* [SMOTE] (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) para obtener una muestra de entrenamiento más equilibrada y así mejorar los resultados de precisión del modelo de clasificación *random forest*. Finalmente, antes de correr los modelos de clasificación se ejecutaron pruebas chi-cuadrado de independencia, para verificar si las variables categóricas empleadas en los clústeres eran independientes entre sí o si mostraban algún tipo de relación.

Herramientas de software

Para el desarrollo de todo este trabajo, fue necesario emplear diferentes herramientas informáticas. Inicialmente, los datos fueron suministrados por RedCB empleando hojas de cálculo en Excel, separadas por mes. Con la finalidad de agruparlos en una base única y poderlos trabajar como series de tiempo, se empleó una base de datos SQL debido al alto número de observaciones: se trabajó con 8'324.029 transacciones ejecutadas en la red, superando los límites asociados a Excel.

Buscando unificar esta base, se eligió trabajar con el programa de base de datos PostgreSQL debido a su poder y su licencia de código abierto. Esta última licencia le permitirá a RedCB el uso de esta base de datos unificada y sus respectivos análisis sin costo adicional, dado que la licencia de código abierto de PostgreSQL estipula su uso gratuito para todo uso.

Siguiendo con la preferencia por el código abierto, se usó R para el análisis de series de tiempo. Las ventajas de R como paquete estadístico consistieron en la facilidad para conectarlo a la base de datos en PostgreSQL sin inconvenientes utilizando el paquete *RPostgres* (Wickham, Ooms, & Müller, 2018), además de la existencia de múltiples paquetes para el cálculo del análisis de clúster en series de tiempo. Ejemplos de lo anterior se pueden ver en los paquetes *jmotif* (Senin, 2018) y *dtwclust* (Sardá-Espinosa, 2017), los cuales fueron empleados para la reducción de dimensiones PAA y el análisis de clúster con distancia DTW respectivamente. Igualmente, se usó el paquete *dendextend* (Galili, 2015) para la creación del árbol asociaado al clúster jerárquico.

Para el ejercicio de minería de texto ejecutado con el fin de determinar la actividad principal de cada comercio, se utilizaron los paquetes *tm* (Feinerer, Hornik, & Meyer, 2008) para la clasificación y el preprocesamiento de texto, *SnowballC* (Bouchet-Valat, 2014) para la extracción de *stems*, *wordcloud2* (Lang & Chien,

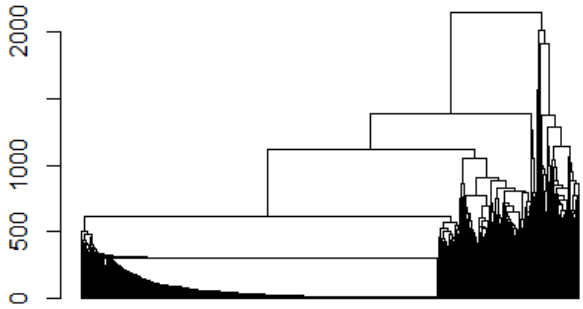
2018) para la nube de términos, *forcats* (Wickham, 2018) para la recodificación y agrupación de los factores asociados la actividad principal de cada comercio, y *DataExplorer* (Cui, 2018) para la visualización final de los niveles de la actividad del comercio.

Igualmente, para la ejecución del SVM se emplearon los paquetes *e1071* (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2018) para la ejecución del SVM, *caTools* (Tuszynski, 2018) para la división de la base con características del comercio entre entrenamiento y prueba, *caret* (Wing etal., 2018) para el cálculo de la matriz de confusión y los indicadores de ajuste del modelo de clasificación, y *pROC* (Robin etal., 2011) para la creación de la curva ROC y el cálculo del AUC. El paquete *randomForest* (Liaw & Wiener, 2002) fue empleado para ejecutar el *random forest*, mientras el paquete *UBL* (Branco, Ribeiro, & Torgo, 2016) se utilizó para ejecutar el remuestreo SMOTE. Otros paquetes empleados fueron los paquetes *ggplot2* (Wickham, 2016), *gridExtra* (Auguie, 2017), *stringi* (Gagolewski, 2018) y *Hmisc* (Jr, Dupont, & others, 2018).

Resultados

Análisis de comportamientos transaccionales de la red a lo largo del tiempo a partir de análisis de clúster

Inicialmente, se ejecutó un análisis de clúster jerárquico, con la finalidad de elegir el número de clústeres a obtener. Dicho análisis jerárquico empleó la distancia DTW sobre la serie reducida con PAA, y su resultado se puede visualizar en la Figura 3.



A partir de los resultados de este dendograma, se planteó particionar los comercios en tres, siete o quince clústeres. Para compararlos, se emplearon indicadores de validación de clústeres (CVI) internos, específicamente los mencionados por Sardá-Espinosa (2017):

- Índice de silueta (Sil), a ser maximizado.
- Función de score (SF), a ser maximizado.
- Índice Calinski-Harabasz (CH), a ser maximizado.
- Índice Davies-Bouldin (DB), a ser minimizado.
- Índice Davies-Bouldin modificado (DBstar), a ser minimizado.
- Índice Dunn (D), a ser maximizado.
- Índice COP (COP), a ser minimizado.

Los resultados para estos CVI internos en las tres particiones propuestas pueden observarse en el Cuadro 4. Tras observar los resultados del Cuadro 4, se eligieron tres particiones para el análisis de clúster. Los resultados de estas particiones se pueden observar en el Cuadro 5.

Cuadro 4: Comparación de resultados para los indicadores de validación de clúster (CVI) internos

	3 particiones	7 particiones	15 particiones
Sil	0.6842114	0.1241122	0.0826562
SF	0.0000000	0.0000000	0.0000000
CH	1193.3101112	552.5613775	280.5819971
DB	1.5656295	7.3516499	5.0773726
DBstar	1.5862310	45.2249154	42.9911928
D	0.0742230	0.0003507	0.0005844
COP	0.1110878	0.0959351	0.0855002

Cuadro 5: Descripción de las particiones elegidas por el análisis de clúster basado en distancia DTW

	Observaciones	Distancia promedio
Clúster 1	96	929.23110
Clúster 2	1748	90.55736
Clúster 3	475	731.79304

Para cada uno de los clústeres, se extrajo la forma de los centroides las series de tiempo asociadas a las alineaciones óptimas obtenidas a partir del análisis de clúster, utilizando la función *shape_extraction* del paquete *dtwclust* (Sardá-Espinosa, 2017). Dicha función se basó en el algoritmo k-Shape propuesto por Paparrizos y Gravano (2015), por lo cual requería normalizar las series de tiempo. Las series obtenidas

para cada uno de los clústeres se pueden observar en la Figura 4.

A partir de los resultados en la Figura 4, se observan comportamientos notoriamente diferentes entre clústeres de corresponsales bancarios. Para los corresponsales del grupo 1, se observó cómo los servicios de información no dejaron de disminuir a lo largo del periodo analizado, mientras los recaudos crecieron hasta alcanzar un pico, y de ahí comenzaron a disminuir. Por otro lado, las recargas y transacciones financieras siguieron un comportamiento similar, con caídas fuertes inicialmente hasta llegar a un punto de inflexión, seguido por un crecimiento leve.

Por otro lado, en el grupo 2 de corresponsales no se vieron diferencias en el comportamiento transaccional según el tipo de transacción. Para todos los cuatro grupos transaccionales, se presentó una caída fuerte en el total de transacciones hasta después de la semana 100, cuando el volumen de transacciones por punto en este grupo volvió a crecer. Esto demuestra un comportamiento irregular, destacando el menor volumen de transacciones.

Finalmente, para el grupo 3 se hallaron dos tipos de comportamiento. Por un lado, las transacciones de recargas y financieras presentaron una caída notoria a lo largo del periodo analizado. Por el otro, dicha caída se vio compensada por un aumento marcado en el volumen de transacciones de recaudo y servicios de información. Desde el punto de vista estratágico, este comportamiento indica que, hasta el momento, RedCB ha tenido problemas para incrementar el volumen de transacciones. Si bien se observó un punto de inflexión en el último año, cuando el volumen de transacciones por punto volvió a crecer para el grupo 2 (el más grande), dicho volumen aún no ha regresado al pico obtenido en 2015, al inicio de la serie.

Por otro lado, cabe recordar que los patrones transaccionales empleados para clasificar los corresponsales bancarios fueron identificados a partir de series normalizadas, por lo cual no son un indicador de los valores absolutos de los corresponsales. Debido a lo anterior, se obtuvo el promedio transaccional de cada uno de los clústeres para agosto de 2018, a partir de la clasificación obtenida y los puntos incluídos en la base de datos. Estos resultados, con los cuales se busca visualizar la escala de los comercios pertenecientes a cada clúster, se pueden observar en el Cuadro 6.

De este, se concluye que los corresponsales en el clúster 1 presentaron el volumen promedio de transacciones más elevado con 658 transacciones mensuanles por punto, seguido por los corresponsales en el clúster 3 con 578 transacciones mensuales por punto. Los corresponsales en el clúster 2 presentaron el promedio transaccional más bajo, con 233 transacciones mensuales por punto. Igualmente, se observó cómo las transacciones financieras presentaron la mayor frecuencia transaccional para todos los grupos, mientras las transacciones de recarga fueron las menos comunes durante dicho mes para todas las categorías.

Cuadro 6: Número de transacciones promedio para agosto de 2018, por clúster transaccional y tipo de transacción

	Clúster 1	Clúster 2	Clúster 3
Financieras	320	108	264
Recargas	38	11	27
Recaudos	226	87	236
Servicios de Información	74	27	50
Total	658	233	578

Tras los resultados obtenidos en el análisis de clúster de series de tiempo, se analizaron los resultados para cada una de las variables descriptivas seleccionadas en la serie. Para esto se utilizó el segundo conjunto de datos, el cual resultó incompleto para lo códigos de los comercios analizados debido a su antigüedad. Por lo anterior, de los 2319 comercios analizados en el análisis de clúster con series de tiempo fue necesario excluir 268 por ausencia de información, llevando a que estas tablas cruzadas fueran ejecutadas únicamente con 2051 corresponsales bancarios. Cabe mencionar que todos los 268 puntos descartados de este análisis pertenecían al grupo 2 del análisis de clúster obtenido, la partición con la mayor cantidad de observaciones.

Así, en el Cuadro 7 se observa el cruce entre el distrito donde se ubica el corresponsal bancario y el clúster bajo el cual fue agrupado. Se observa cómo la mayor parte de los puntos del grupo 1 se encuentra en el distrito de Ate, mientras San Juan de Lurigancho es el distrito con mayor participación en el grupo 2. Lo anterior parece dar cuenta de diferencias en la distribución de los distritos entre clústeres, lo cual se corroboró en etapas posteriores del presente documento.

Por otro lado, se obtuvo un ingreso familiar per cápita ponderado para cada uno de los clústeres a partir de la frecuencia de corresponsales bancarios en un clúster y la información para el ingreso familiar

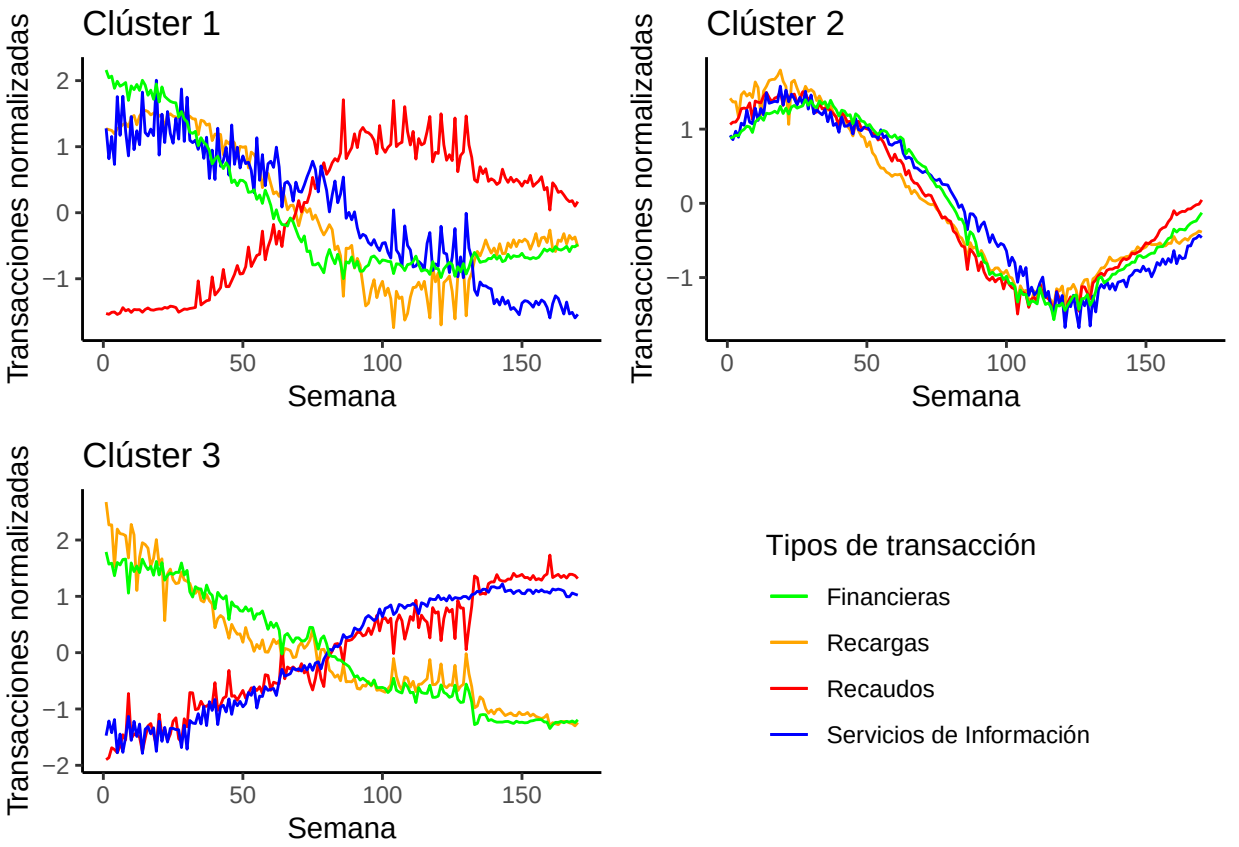


Figura 3: Resultados del análisis de clúster particional

per cápita por distrito en Perú a 2012, cuya fuente fue una base de datos suministrada por el Programa de las Naciones Unidas para el Desarrollo (PNUD, 2012). Se obtuvo que el ingreso familiar promedio para el clúster 1 era PEN 1007, mientras el ingreso familiar promedio para el clúster 2 fue PEN 1045. Finalmente, para el clúster 3, el ingreso familiar promedio fue 1023. De lo anterior se concluye que no existen diferencias significativas en cuanto al nivel socioeconómico de los clústeres.

Finalmente, el Cuadro 9 cruza el clúster asignado a los corresponsales con la sitaución actual del mismo. Se observa cómo la mayoría de los puntos deshabilitados en la actualidad alguna vez fueron corresponsales del grupo 2, mientras la mayor parte de los corresponsales en los grupos 1 y 3 aún opera en la red.

Verificación de relaciones entre los clústeres transaccionales y características de los corresponsales bancarios

Tras la obtención de las categorías de corresponsales bancarios, fue necesario verificar la existencia de relaciones entre estos clústeres y variables categóricas asociadas a los corresponsales bancarios, así como la existencia de relaciones entre estas y el estado actual de los mismos. Para ello, se empleó la técnica chi-cuadrado de doble vía sobre cuatro de estas relaciones. Así, el Cuadro 10 permite observar que existe una relación significativa entre el distrito del corresponsal y el clúster al cual este pertenece. Lo anterior confirma lo observado en el Cuadro 7, por lo cual se concluye que el distrito donde se encuentra ubicado el corresponsal es un criterio relacionado con la futura transaccionalidad del mismo, y debe ser tenido en cuenta al momento de seleccionar puntos en el futuro.

A su vez, el Cuadro 11 demuestra que no existe

Cuadro 7: Tabla cruzada entre el distrito del corresponsal bancario y el clúster asignado al mismo

	Clúster 1	Clúster 2	Clúster 3
Ancon	0	1	1
Ate	15	83	18
Barranco	0	4	1
Bellavista	2	33	4
Breña	4	33	6
Callao	0	62	26
Carabayllo	1	27	20
Carmen De La Legua Reynoso	3	13	6
Chaclacayo	0	10	3
Chorrillos	0	25	7
Cieneguilla	0	3	1
Comas	7	33	23
El Agustino	1	46	6
Independencia	2	19	13
Jesus Maria	0	35	10
La Molina	0	6	6
La Perla	2	21	4
La Punta	0	2	0
La Victoria	1	57	17
Lima	7	132	35
Lince	1	17	5
Los Olivos	5	59	32
Lurigancho	2	43	6
Lurin	1	9	3
Magdalena Del Mar	0	4	0
Miraflores	0	3	0
Pachacamac	1	9	1
Pucusana	0	3	2
Pueblo Libre	0	12	6
Puente Piedra	5	31	15
Punta Hermosa	0	2	1
Punta Negra	0	1	0
Rimac	0	22	8
San Bartolo	0	1	0
San Borja	0	6	1
San Isidro	0	1	0
San Juan De Lurigancho	4	135	37
San Juan De Miraflores	3	66	18
San Luis	0	10	4
San Martin De Porres	2	90	40
San Miguel	1	36	5
Santa Anita	5	77	14
Santiago De Surco	1	9	6
Surquillo	3	10	2
Ventanilla	3	27	17
Villa El Salvador	12	78	27
Villa Maria Del Triunfo	2	74	18

Cuadro 8: Tabla cruzada entre el distrito del corresponsal bancario y el estado actual del mismo

	Activo	Deshabilitado
Ancon	1	1
Ate	26	90
Barranco	1	4
Bellavista	3	36
Breña	6	37
Callao	18	70
Carabayllo	15	33
Carmen De La Legua Reynoso	8	14
Chaclacayo	4	9
Chorrillos	6	26
Cieneguilla	1	3
Comas	23	40
El Agustino	7	46
Independencia	12	22
Jesus Maria	7	38
La Molina	4	8
La Perla	4	23
La Punta	0	2
La Victoria	16	59
Lima	35	139
Lince	4	19
Los Olivos	23	73
Lurigancho	5	46
Lurin	3	10
Magdalena Del Mar	0	4
Miraflores	0	3
Pachacamac	4	7
Pucusana	2	3
Pueblo Libre	5	13
Puente Piedra	14	37
Punta Hermosa	1	2
Punta Negra	0	1
Rimac	8	22
San Bartolo	0	1
San Borja	1	6
San Isidro	0	1
San Juan De Lurigancho	33	143
San Juan De Miraflores	25	62
San Luis	4	10
San Martin De Porres	29	103
San Miguel	7	35
Santa Anita	19	77
Santiago De Surco	8	8
Surquillo	4	11
Ventanilla	11	36
Villa El Salvador	33	84
Villa Maria Del Triunfo	17	77

ninguna relación entre la actividad principal del corresponsal y el clúster transaccional dentro del cual fue clasificado.

Por otro lado, el Cuadro 12 muestra que no existe una relación significativa entre el distrito del corresponsal y su estado actual, al observarse un p-valor

Cuadro 9: Tabla cruzada entre el clúster asignado al corresponsal bancario y el estado actual del mismo

	Activo	Deshabilitado
Clúster 1	82	14
Clúster 2	48	1432
Clúster 3	327	148

Cuadro 10: Prueba chi-cuadrado entre el distrito del corresponsal bancario y el clúster asignado al mismo

Chi-cuadrado	Grados de libertad	P-valor
157.3266	92	2.65e-05

Cuadro 11: Prueba chi-cuadrado entre la actividad principal del corresponsal bancario y el clúster asignado al mismo

Chi-cuadrado	Grados de libertad	P-valor
21.40177	16	0.1635789

superior a 0.05 para el estadístico de prueba. De lo anterior se concluye que estas variables son independientes.

Cuadro 12: Prueba chi-cuadrado entre el distrito del corresponsal bancario y el estado actual del mismo

Chi-cuadrado	Grados de libertad	P-valor
57.73166	46	0.1149709

Finalmente, el Cuadro 13 demuestra que existe una relación significativa entre el clúster asociado al corresponsal bancario y su estado actual, al observarse un p-valor inferior a 0.05.

Cuadro 13: Prueba chi-cuadrado entre el clúster asignado al corresponsal bancario y el estado actual del mismo

Chi-cuadrado	Grados de libertad	P-valor
1125.391	2	0

Determinación de la actividad principal del comercio por medio de minería de texto

Para esta sección fue necesario preprocesar la variable elegida para la extraccion de la actividad principal, específicamente el nombre del comercio que ejerce como corresponsal bancario. Durante dicho preprocesamiento se convirtieron todos los caracteres a minúscula, se eliminaron puntos, números y dobles espacios, y se eliminaron *stop words* comunes al idioma español. Finalmente, se extrajeron *stems* a partir de las palabras preprocesadas.

Con lo anterior, se obtuvo una matriz para la frecuencia de cada uno de los *stems* hallados. De estos, el Cuadro 14 muestra los 10 *stems* más frecuentes.

Cuadro 14: Tabla de frecuencia para las diez palabras más comunes en el nombre del comercio

	Palabra	Frecuencia
bodega	bodega	410
botica	botica	267
bazar	bazar	216
libreria	libreria	202
comerci	comerci	64
farma	farma	62
minimarket	minimarket	59
multiservicio	multiservicio	52
locutorio	locutorio	45
internet	internet	40

A continuación, se creó una nueva variable denominada “Tipo”, con la palabra en el nombre del comercio más común a lo largo del conjunto de datos. Por ejemplo, para un comercio llamado “bodega Magaly”, el valor de la variable “Tipo” sería “Bodega”. Tras un preprocesamiento adicional se obtuvo dicha columna, la cual fue agregada a la base con características descriptivas de cada corresponsal bancario, y convertida en un factor.

Por otro lado, tras visualizar los niveles de los factores se observaron niveles con nombres similares, como “bodega” y “bodeguita”. Dado lo anterior, se agruparon los siguientes niveles de “Tipo”:

- libreria_bazar: incluye los niveles “libreria”, “bazar”, “bazarlibreria”, “bazarlicoreria”, “licoreria”.
- botica: incluye los niveles “farma”, “boticfarma”,

"farmacia", "salud", "botica", "biofarma", "lafarma".
- bodega: incluye los niveles "bodega", "bodeguita", "bodegon", "bogeda", "tienda".
- cybercafe: incluye los niveles "ciberequi", "cibernet", "ciber", "cyber", "internet", "computel", "cybermrket", "cybergam", "cabina".
- locutorio: incluye los niveles "locutorio", "telecomunicacion", "comunicacion".
- multiservicios: incluye los niveles "multiservic", "multiservicio", "multipago", "servic", "servicio".
- minimercado: incluye los niveles "minimarket", "mini", "market", "supermarket".

Figura 4: Nube de palabras con los términos más frecuentes en la base de comercios, antes de la extracción de *stems*

Finalmente, se mantuvieron únicamente los niveles con una participación superior al 2 % en el total de los comercios, agrupando todos los otros niveles bajo la categoría "otros". Esto se debe a que, debido a los nombres de los comercios, no fue posible determinar con claridad su actividad principal. Igualmente, múltiples actividades tenían frecuencias muy bajas en la red, como "panaderia" y "ferreteria". La tabla de frecuencias final puede ser observada en el Cuadro 15.

Cuadro 15: Tabla de frecuencias final para la actividad principal de los corresponsales bancarios

Tipo	Conteo	Porcentaje	Acumulado
otros	752	0.3666504	0.3666504
bodega	419	0.2042906	0.5709410
botica	312	0.1521209	0.7230619
libreria_bazar	252	0.1228669	0.8459288
multiservicios	76	0.0370551	0.8829839
minimercado	75	0.0365675	0.9195514
comerci	63	0.0307167	0.9502682
locutorio	54	0.0263286	0.9765968
cybercafe	48	0.0234032	1.0000000

Finalmente, las Figuras 5, 6, 7 y 8 muestran las nubes para las tablas de frecuencia finales, tanto para todos los comercios como para cada uno de los clúster. No se aprecian diferencias visuales fuertes en los tamaños de las palabras entre clústeres, permitiendo visualizar los resultados obtenidos en la prueba chi-cuadrado.

Figura 5: Nube de palabras con las actividades en los comercios del clúster 1

Figura 6: Nube de palabras con las actividades en los comercios del clúster 2

Figura 7: Nube de palabras con las actividades en los comercios del clúster 3

Clasificación del estado del corresponsal a partir de su comportamiento transaccional

Tras lo anterior, fue necesario clasificar el estado del corresponsal bancario a partir de su comportamiento transaccional en el pasado, para lo cual se empleó el modelo SVM, que debía ser especificado. Se determinó una fórmula bajo la cual la variable "Estado" actuaría como la variable a clasificar, con la variable "Cluster" actuando como la variable clasificadora. Nuevamente, se creó una variable dicotómica por tipo de clúster, obteniendo tres variables en total. Se obtuvo que la mejor forma funcional era un modelo SVM con costo 0.5, gamma 0.25 y kernel polinomial. De este modelo se obtuvieron 632 vectores de soporte. Todo lo anterior se ejecutó sobre la base de entrenamiento obtenida.

Tras lo anterior, se predijeron los valores del estado actual del corresponsal para la base de prueba. Dichos valores fueron comparados con los originales en una matriz de confusión, introducida en el Cuadro 16. Los indicadores de ajuste generales y específicos fueron incorporados en los Cuadros 17 y 18.

Cuadro 16: Matriz de confusión para el modelo SVM

	Activo	Deshabilitado
Activo	110	34
Deshabilitado	13	356

Cuadro 17: Indicadores de ajuste generales para el modelo SVM

	x
Accuracy	0.9083821
Kappa	0.7625631
AccuracyLower	0.8800312
AccuracyUpper	0.9319057
AccuracyNull	0.7602339
AccuracyPValue	0.0000000
McnemarPValue	0.0035308

Cuadro 18: Indicadores de ajuste específicos para el modelo SVM

	x
Sensitivity	0.8943089
Specificity	0.9128205
Pos Pred Value	0.7638889
Neg Pred Value	0.9647696
Precision	0.7638889
Recall	0.8943089
F1	0.8239700
Prevalence	0.2397661
Detection Rate	0.2144250
Detection Prevalence	0.2807018
Balanced Accuracy	0.9035647

Finalmente, se obtuvo la curva ROC. Dicha curva puede visualizarse en la Figura 9. El AUC asociado

se puede apreciar en el Cuadro 19.

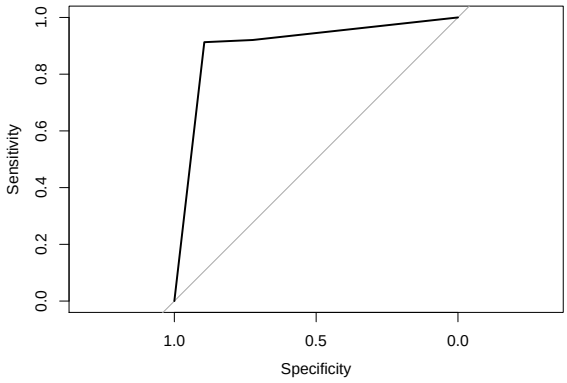


Figura 8: Curva ROC

Cuadro 19: AUC para el modelo de clasificación SVM

x
0.8995622

A partir de lo anterior, se pudo apreciar que el clúster transaccional fue un predictor preciso para el estado actual del punto, tanto por AUC como por precisión. Estos resultados corroboran los resultados de la prueba chi-cuadrado, al tiempo que confirma la idea de que el comportamiento de las transacciones a lo largo del tiempo afecta directamente el cierre o no de un corresponsal. Lo anterior constituye un hallazgo para las necesidades de RedCB, ya que le permitirá seguir mejor a sus puntos actuales y vigilar cuáles están en riesgo de cerrar.

La relevancia de este hallazgo se debe a que, hoy, RedCB tiene como único criterio de selección para cierre de puntos el que estos no alcancen un número mínimo de transacciones durante cierto periodo de tiempo. Lo anterior, si bien sirve para eliminar puntos no rentables, no permite identificar puntos con potencial de repetir este patrón en el futuro, impidiendo medidas correctivas y limitándose a reacciones. A partir de lo anterior, RedCB puede enfocarse en estos puntos, identificar sus características comunes, y formular estrategias para lograr aumentar la transaccionalidad de los mismos, con la finalidad de evitar el cierre de estos.

Identificación y clasificación de comportamientos transaccionales a partir de características del comercio

En esta sección, se clasificaron los corresponsales en uno de los clústeres transaccionales hallados inicialmente a partir del distrito donde están ubicados y su actividad principal. Con el fin de clasificar, se eligió ejecutar un modelo *random forest* sobre la base final con las características de los comercios que pertenecen o alguna vez pertenecieron a la red de RedCB. Dicha base, además de los valores originales suministrados por RedCB, incluye los valores obtenidos a partir de la minería de texto para la actividad principal de cada comercio.

Con el fin de reducir dimensiones, se eliminaron categorías con menos de 30 observaciones. Esto llevó a la eliminación de 24 variables dicotómicas, todas ellas distritos, por lo cual se borró la información sobre el distrito de 252 puntos, los cuales quedarían con valores cero para todas las dicotómicas de distrito que permanecieron. La base fue dividida entre una base de entrenamiento y una base de pruebas, a una tasa 75/25.

Sobre la base de entrenamiento obtenida, se ejecutó la técnica de remuestreo SMOTE. Las categorías para la muestra rebalanceada pueden ser apreciadas en el Cuadro 20.

Cuadro 20: Distribución por categorías de transacciones a lo largo del tiempo tras remuestreo SMOTE

Var1	Freq
1	1633
2	1654
3	1629

Sobre dicha base remuestreada, se estimó el modelo de *random forest*. La matriz de confusión obtenida tras estimar valores para el clúster sobre el conjunto de prueba puede ser observada en el Cuadro 21. Igualmente, los indicadores de ajuste generales y específicos se aprecian en los Cuadros 22 y 23. El AUC para este modelo se observa en el Cuadro 24. Cabe mencionar que se utilizó la implementación de ROC multiclase existente en el paquete *pROC*, basada en la técnica propuesta por Hand y Till (2001).

Finalmente, se calculó la importancia de cada variable en la clasificación obtenida a partir del modelo

Cuadro 21: Matriz de confusión para el modelo *random forest*

1	2	3
9	69	22
7	179	45
9	129	46

Cuadro 22: Indicadores de ajuste generales para el modelo *random forest*

	x
Accuracy	0.4543689
Kappa	0.0654263
AccuracyLower	0.4107643
AccuracyUpper	0.4985002
AccuracyNull	0.7320388
AccuracyPValue	1.0000000
McnemarPValue	0.0000000

Cuadro 23: Indicadores de ajuste específicos para el modelo *random forest*

	Class: 1	Class: 2	Class: 3
Sensitivity	0.3600000	0.4748011	0.4070796
Specificity	0.8142857	0.6231884	0.6567164
Pos Pred Value	0.0900000	0.7748918	0.2500000
Neg Pred Value	0.9614458	0.3028169	0.7975831
Precision	0.0900000	0.7748918	0.2500000
Recall	0.3600000	0.4748011	0.4070796
F1	0.1440000	0.5888158	0.3097643
Prevalence	0.0485437	0.7320388	0.2194175
Detection Rate	0.0174757	0.3475728	0.0893204
Detection Prevalence	0.1941748	0.4485437	0.3572816
Balanced Accuracy	0.5871429	0.5489947	0.5318980

Cuadro 24: AUC para el modelo de clasificación *random forest*

x
0.5471895

random forest, el cual se observa en la Figura 10.

A partir de lo anterior, se observaron resultados regulares para el modelo de clasificación, tanto por AUC como por precisión. Lo anterior permite concluir que se necesitan más variables además del distrito y el tipo de comercio para predecir efectivamente el clúster del comercio. Aún así, el hecho que exista una correlación y cierta precisión demuestran tanto las ventajas del remuestreo SMOTE como el uso del *random forest* como modelo de clasificación.

Igualmente, se apreció que el tipo de comercio tenía un alto grado de influencia sobre los resultados del modelo. En la Figura 10, se pudo apreciar cómo tres de las primeras cinco variables más influyentes en la precisión eran tipos de comercio, siendo estos cuatro de las cinco variables más influyentes sobre el índice Gini. Lo anterior indica que, al momento de asignar un clúster transaccional a los comercios, el tipo de comercio es una variable significativa. Aún así, el distrito del comercio sigue siendo un conjunto de factores relevante, que podría ser complementado a futuro con variables geoespaciales adicionales.

Conclusiones

Por un lado, el modelo de minería de texto permitió obtener resultados satisfactorios para el tipo de comercio. Sin embargo, hubo limitantes debido a la falta de información para el tipo de comercio por parte de RedCB. Lo anterior impidió la creación de un modelo de clasificación de minería de texto a partir de los nombres, lo cual se puede constituir en una línea futura de estudio una vez se obtengan los datos requeridos.

Igualmente, las pruebas chi-cuadrado de dos vías ejecutadas permitieron constatar la existencia de relaciones significativas entre el clúster transaccional y distrito donde se ubica el clúster, indicando una dimensión geográfica en el comportamiento transaccional. También se halló una relación significativa entre el clúster transaccional y el estado actual del corresponsal, por lo cual se reafirma la concepción de que la permanencia de un corresponsal bancario depende de su evolución transaccional en el tiempo y el volumen transaccional alcanzado por estos.

A partir del análisis de clúster para los comportamientos transaccionales de los puntos a lo largo del tiempo se identificó cómo RedCB aún no ha terminado de recuperarse de la caída en el volumen de transacciones debido a su política de expansión indiscriminada entre 2015 y 2017. Esto se ve tanto en el alto número de corresponsales con baja transaccionalidad (el promedio transaccional del clúster 2 fue

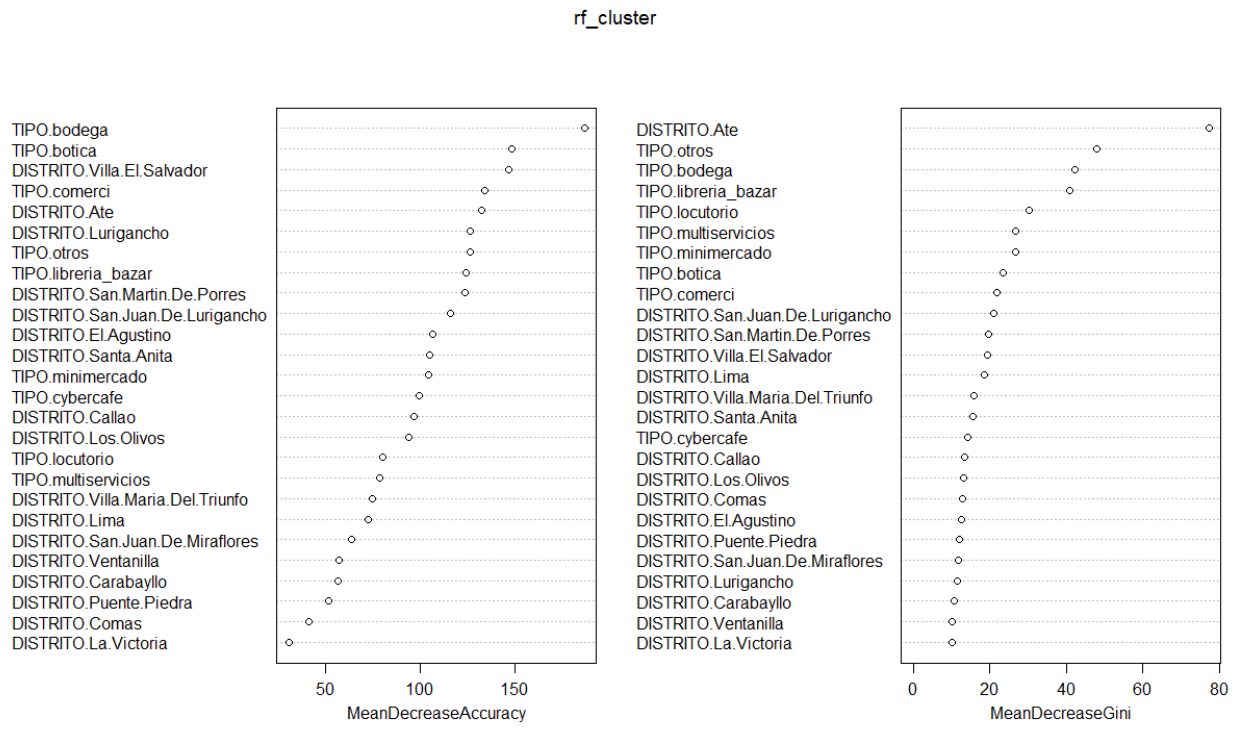


Figura 9: Importancia de las variables en el modelo *random forest*

233 transacciones mensuales por punto) y el repunte reciente en los totales transaccionales en la Figura 4, el cual aún está en sus inicios.

A su vez, se apreció cómo en ciertos comercios las transacciones de recaudos y servicios de información han crecido, a costa de las recargas y los servicios financieros. Lo anterior presenta ventajas y retos para RedCB, debido a la estructura de remuneración asociada a los servicios de corresponsalía bancaria, bajo la cual los bancos para los cuales opera le pagan según el tipo de transacción ejecutado, siendo las recargas el tipo de transacción para el cual RedCB recibe una menor remuneración. Sin embargo, la caída en los servicios financieros tampoco ayuda a la rentabilidad de los puntos, dado que es el tipo de transacción que más paga. De lo anterior se concluye que RedCB necesita potenciar este tipo de transacciones en sus puntos. A petición de RedCB, la estructura de remuneración se mantuvo como confidencial.

Por otro lado, el modelo de clasificación para los clústeres y el estado del corresponsal arrojó buenos resultados tanto para el AUC como para la precisión. Con lo anterior se concluye que la transaccionalidad actúa como predictor para la persistencia o el cierre de corresponsales bancarios. Esto puede ser utilizado como insumo para la toma de decisiones de apertura por parte de RedCB, ya que podría a partir de las características de los comercios potenciales clasificarlos en un clúster transaccional, y a partir de estos decidir sobre la apertura o no teniendo en cuenta su estado esperado en el futuro. Para lo anterior se recomienda complementar los resultados con investigación futura, perfeccionando el modelo de clasificación y analizando la posibilidad de un modelo de regresión.

En cuanto a la clasificación de clústeres según características descriptivas empleando un modelo *random forest*, no se obtuvieron los resultados esperados en cuanto a la precisión del modelo, potencialmente debido a un problema de variables omitidas. Específicamente, se podría sugerir el uso de ubicaciones exactas

para los puntos de la red, así como más información acerca de los niveles de ventas de los puntos en sus operaciones principales. Lo anterior puede sugerirse como tema de investigación futura.

A partir de lo anterior, se observó cómo el tipo de comercio pesa más que el distrito donde este se encuentra ubicado como criterio de clasificación para la transaccionalidad futura de los mismos. Lo anterior, sumado al hecho que no habían diferencias importantes en el ingreso familiar per cápita promedio para los clústeres transaccionales, lleva a concluir que RedCB debe enfatizar más en la actividad de los mismos al seleccionar nuevos puntos. En el futuro, un análisis sobre los montos transados podría convertirse en un complemento para emitir recomendaciones estratégicas acerca del tipo de comercio y los distritos en los cuales RedCB se debe enfocar al reclutar nuevos corresponsales bancarios.

El presente estudio agrega valor a RedCB por medio de un mejor conocimiento de su red actual de puntos y su permanencia, a partir del análisis a lo largo del tiempo. De lo anterior se observaron tres patrones transaccionales que permitieron identificar problemas en la operación, como una correlación inversa en algunos corresponsales entre recaudos y servicios financieros, los dos tipos de transacción más rentables para RedCB. Lo siguiente consiste en identificar los motivos detrás de este patrón, para lo cual se espera que el presente estudio sea un insumo.

Finalmente, se observó cómo tanto el tipo de comercio como la ubicación del corresponsal influyeron sobre el clúster transaccional al cual pertenecían. Nuevamente, esta información es una herramienta para optimizar la toma de decisiones de apertura por parte de RedCB, al permitirle afinar sus criterios de selección al enfocarse en tipos de comercio con mayor rentabilidad o mejores patrones de comportamiento. Esto va ligado a la necesidad de corregir los patrones transaccionales que muestran esa correlación inversa entre recaudos y servicios financieros.

Referencias

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for «Grid» Graphics*. Recuperado de https://CRAN.R-project.org/package=gridExtra

Bouchet-Valat, M. (2014). *SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library*. Recuperado de https://CRAN.R-project.org/package=SnowballC

Branco, P., Ribeiro, R. P., & Torgo, L. (2016). UBL: an R Package for Utility-Based Learning. *CoRR*, *abs/1604.08079*.

Cardona Prada, J. C. (2017). *Location-allocation problem for banking correspondent services: the colombian urban market case* (PhD Thesis). Pontificia Universidad Católica del Perú, CENTRUM; Maastricht School of Management.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Cui, B. (2018). *DataExplorer: Data Explorer*. Recuperado de https://CRAN.R-project.org/package=DataExplorer

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*, *25*(5), 1-54. Recuperado de http://www.jstatsoft.org/v25/i05/

Gagolewski, M. (2018). *R package stringi: Character string processing facilities*. Recuperado de http://www.gagolewski.com/software/stringi/

Galili, T. (2015). dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btv428

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., Tatham, R. L., & others. (2006). *Multivariate data analysis (Vol. 6)*. Upper Saddle River, NJ: Pearson Prentice Hall.

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine learning*, *45*(2), 171-186.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Jayo, M. (2010). *Correspondentes bancários como canal de distribuição de serviços financeiros: taxonomia, histórico, limites e pontencialidades dos modelos de gestão de redes* (PhD Thesis). Fundação Getulio Vargas.

Jr, F. E. H., Dupont, with contributions from C., & others, many. (2018). *Hmisc: Harrell Miscella-*

neous. Recuperado de https://CRAN.R-project.org/package=Hmisc

Keogh, E. J., & Pazzani, M. J. (2000). Scaling up dynamic time warping for datamining applications. ACM. Recuperado de https://dl.acm.org/citation.cfm?id=347153

Keogh, E., Chakrabarti, K., Pazzani, M., & Mehrotra, S. (2001). Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and information Systems*, *3*(3), 263-286.

Lang, D., & Chien, G.-t. (2018). *wordcloud2: Create Word Cloud by 'htmlwidget'*. Recuperado de https://CRAN.R-project.org/package=wordcloud2

Liao, T. W. (2005). Clustering of time series dataa survey. *Pattern recognition*, *38*(11), 1857-1874.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18-22. Recuperado de https://CRAN.R-project.org/doc/Rnews/

Loureiro, E., Madeira, G., & Bader, F. (2011). Expansão dos correspondentes bancários no Brasil: uma análise empírica. *Texto para discussão*, (433).

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2018). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. Recuperado de https://CRAN.R-project.org/package=e1071

Paparrizos, J., & Gravano, L. (2015). k-shape: Efficient and accurate clustering of time series. ACM. Recuperado de www1.cs.columbia.edu/~jopa/Papers/PaparrizosSIGMOD2015.pdf

PNUD, P. de las N. U. para el D. (2012). Índice de Desarrollo Humano departamental, provincial y distrital 2012. Recuperado de http://www.pe.undp.org/content/dam/peru/docs/Publicaciones%20pobreza/INDH2013/pe.Indice%20de%20Desarrollo%20Humano%20Per%C3%BA.xlsx

Ramakrishnan, K. (2010). The competitive response of small, independent retailers to organized retail: Study in an emerging economy. *Journal of Retailing and Consumer Services*, *17*(4), 251-258.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77.

Roelofsen, P. (2018). *Time series clustering* (Tesis de licenciatura). Vrije Universiteit Amsterdam. Recuperado de https://beta.vu.nl/nl/Images/stageverslag-roelofsen_tcm235-882304.pdf

Ryan, J. A., & Ulrich, J. M. (2018). *xts: eXtensible Time Series*. Recuperado de https://CRAN.R-project.org/package=xts

Sardá-Espinosa, A. (2017). Comparing time-series clustering algorithms in r using the dtwclust package. *R package vignette*, *12*.

Senin, P. (2018). *jmotif: Time Series Analysis Toolkit Based on Symbolic Aggregate Dicretization, i.e. SAX*. Recuperado de https://CRAN.R-project.org/package=jmotif

Sørensen, C. K., & Puigvert Gutiérrez, J. M. (2006). *Euro area banking sector integration: using hierarchical cluster analysis techniques*. ECB working paper.

Tuszynski, J. (2018). *caTools: Tools: moving window statistics, GIF, Base64, ROC AUC, etc.* Recuperado de https://CRAN.R-project.org/package=caTools

Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software*, *21*(12), 1-20. Recuperado de http://www.jstatsoft.org/v21/i12/

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Recuperado de http://ggplot2.org

Wickham, H. (2018). *forcats: Tools for Working with Categorical Variables (Factors)*. Recuperado de https://CRAN.R-project.org/package=forcats

Wickham, H., François, R., Henry, L., & Müller, K. (2018). *dplyr: A Grammar of Data Manipulation*. Recuperado de https://CRAN.R-project.org/package=dplyr

Wickham, H., Ooms, J., & Müller, K. (2018). *RPostgres: 'Rcpp' Interface to 'PostgreSQL'*. Recuperado de https://CRAN.R-project.org/package=RPostgres

Wing, M. K. C. from J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., ... Hunt, T. (2018). *caret: Classification and Regression Training*. Recuperado de https://CRAN.R-project.org/package=caret