INFORME DATATON BANCOLOMBIA 2018

Grupo Andinolytics

1. Introducción

En el año 2017 se realizaron más de 52 millones de transacciones electrónicas en Colombia, a través de Pagos Seguros en Línea (PSE), las cuales movilizaron $128.163 millones de cuentas de ahorro y corrientes (Revista Dinero, 2 de agosto de 2018). A diferencia de las transacciones realizadas en establecimientos comerciales mediante un datáfono, las cuales tienen asociado un código MCC (Merchant Category Code) que permite clasificarlas de acuerdo con el tipo de bien o servicio, las transacciones PSE no tienen categorías definidas. En este contexto se enmarca el reto de la competencia Dataton BC 2018, el cual tiene como objetivo principal la categorización de dichas transacciones por medio de técnicas de analítica. Bancolombia tiene a su disposición información de las transacciones PSE y datos relevantes de sus clientes, con los cuales el grupo Andinolytics se propone abordar este desafío y proponer una metodología de analítica para la categorización de las transacciones PSE.

Adicionalmente, los resultados de la categorización se pueden vincular a las apps que administran las finanzas personales o PFM, para añadir nuevas funcionalidades que les faciliten a los usuarios llevar el control de su dinero.

2. Objetivos

- Proponer una metodología para la categorización de las transacciones PSE.
- Definir categorías para las transacciones PSE de acuerdo con el tipo de bien o servicio.
- Desarrollar un modelo para categorizar las transacciones PSE.

3. Metodología

Para abordar el proyecto, se siguieron los siguientes pasos:

A. Entendimiento del problema de negocio.
B. Revisión y limpieza de la información de transacciones y clientes.
C. Análisis lexicográfico sobre transacciones sectorizadas.
D. Definición de categorías de interés de las transacciones PSE.
E. Clasificación de un grupo de transacciones basada en las referencias de recaudadores.
F. Construcción de modelos para la clasificación de transacciones categorizadas.
G. Evaluación de modelos de clasificación.
H. Clasificación de transacciones sin categorizar mediante el mejor clasificador encontrado.
I. Aplicación de categorización en el contexto de PFMs.

4. Descripción y preparación de los datos

Los organizadores de la competencia suministraron a los equipos con dos bases de datos que contienen una muestra de los clientes y las transacciones PSE realizadas entre 2016 y 2018 (Figura 1).

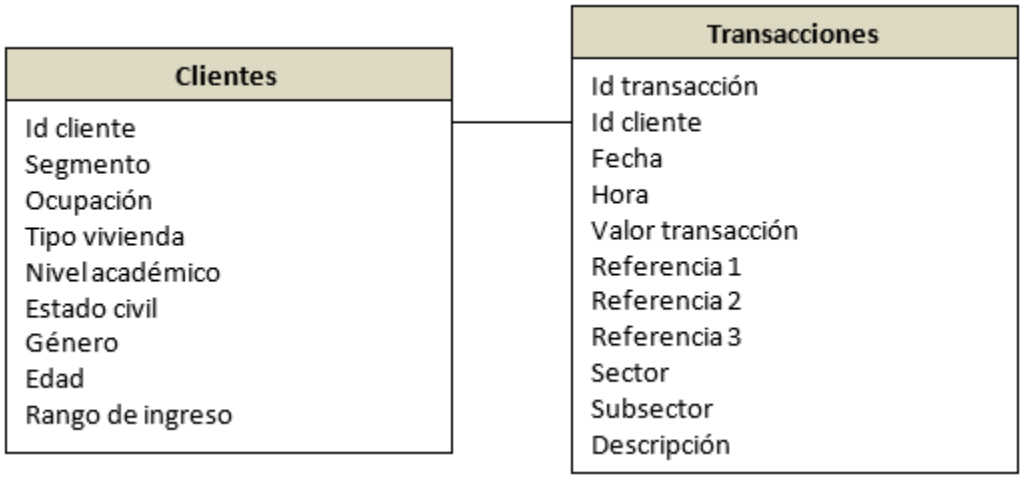

Figura 1. Descripción de las bases de datos

Para la limpieza de la base de datos de transacciones, en primer lugar, se eliminaron aquellos registros con más de 11 comas. Esto dado que los campos de texto libre (Referencia 1, Referencia 2 y Referencia 3) podían contener más comas que dificultaban la separación de los registros en 11 variables. Posteriormente, se eliminaron los registros con valores nulos en el valor de la transacción, al igual que valores inferiores a 1000 COP y superiores a 50 millones COP, los cuales corresponden a menos del 1% de la información. En la Tabla 1 se presenta un resumen de la variable Valor transacción después de la limpieza.

Variable	Observaciones	Media	Desviación	Mínimo	Máximo
Valor transacción	11’809.972	360.800	975.940,9	1.000	49’950.000

Tabla 1. Estadísticas descriptivas de Valor transacción

Adicionalmente, se crearon las siguientes variables categóricas que almacenan información valiosa de la fecha y hora de la transacción:

- Mes: Variable con 12 categorías (una por cada mes del año).
- Dia semana: Variable con 7 categoría (una por cada día de la semana).
- Fin de semana: Variable con 2 categorías (fin de semana y entre semana).
- Quincena: Variable con 3 categorías (primera quincena, segunda quincena y no quincena). La primera quincena va el 1 al 6 de cada mes, y la segunda del 14 al 19.
- Momento del día: Variable con 5 categorías (madrugada, mañana, mediodía, tarde y noche). Madrugada comprende el rango de 2:00 a 6:00, mañana de 6:00 a 11:00, mediodía de 11:00 a 14:00, tarde de 14:00 a 19:00 y noche de 19:00 a 2:00.

Con relación a la base de clientes, se asignaron los valores faltantes a la categoría “I” (no informa) y se creó una variable categórica de edad (menores de 25, 25-30, 31-35, 36-40, 41-45, 46-50, 51-59, y mayores de 60).

5. Definición de las Categorías

Para la definición de las categorías se desarrolló un algoritmo lexicográfico que separaba las palabras de los campos Referencia 1, Referencia 2 y Referencia 3, las cuales contienen la información suministrada por los recaudadores acerca de la transacción PSE realizada. Luego, estas eran almacenadas en un diccionario y se realizaba un conteo. Para lo anterior se leyeron todas las palabras en mayúsculas y se eliminaron las tildes. Posteriormente, se hizo una revisión de las palabras con una frecuencia superior a mil y se asignaron categorías de acuerdo con el tipo de bien o servicio. Es importante mencionar que a conectores (“de”, “y”, “para”, etc.) o palabras con significado ambiguo (“pago”, “valor”, “compra”, etc.) no se les asignó ninguna categoría. Como resultado, se lograron clasificar 273 palabras en 17 categorías (Tabla 2).

Categorías	Palabras	Ejemplos
Banco	48	Colpatria, Citibank, Davivienda
E-commerce	35	Éxito, Alkosto, Ktronix
Educación	31	Colegio, universidad, curso
Servicios	26	Gas, acueducto, Codensa
Telecomunicaciones	20	Internet, móviles, Claro
Entretenimiento	18	Natación, futbol, Cine
Tiquetes	16	Avianca, flight, vuelos
Seguros	12	Suramericana, AXA, Allianz
Trámites y documentos	11	Certificados, trámites, apostilla
Salud	10	Colsanitas, medicina, Coomeva
Alimentación	10	Alimentos, restaurante, almuerzo
Vehículos	9	Vehículos, automotores, vehicular
Hoteles y viajes	7	Booking, turístico, Decamerón
Impuestos	6	Predial, IUVA, tributario
Vivienda	6	Arrendamiento, vivienda, inmobiliaria
Pensiones y cesantías	5	Pensión, pensiones, cesantías
Transporte	3	Transporte, PaseYa, UBER

Tabla 2. Categorías de las transacciones

Posteriormente, se desarrolló un algoritmo para clasificar las transacciones en cada una de las categorías definidas previamente. Para esto, se hizo una unión de los campos Referencia 1, Referencia 2 y Referencia 3 con el carácter espacio. Luego, sobre la cadena resultante se extrajeron las palabras en una lista. Por cada palabra se busca en el diccionario de categorías y se asigna a cuál pertenece. En caso de que hubiese varias clasificaciones, el algoritmo escoge una basado en la categoría que más veces ha sido asignada. Si existe un empate, como heurística se elige la categoría que aparece en primer lugar, con la intuición de que la información más importante se encuentra en la Referencia 1. Finalmente, se categorizaron 3’068.194 transacciones (Figura 2).

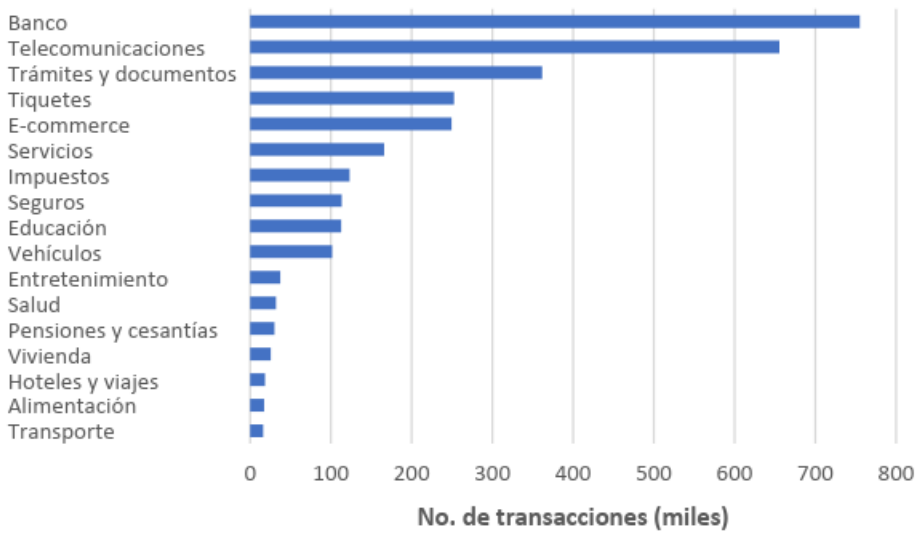

Figura 2. Número de transacciones por categoría

6. Modelo de clasificación

Los problemas de clasificación de datos son ampliamente estudiados en el campo de análisis de información. La clasificación se divide principalmente en supervisada y no supervisada. Los primeros corresponden a algoritmos que basados en un conjunto de datos clasificados (valores de entrenamiento), asignan una clasificación a un segundo conjunto de datos sin clasificar. Por otro lado, en los sistemas de clasificación no supervisados no se cuenta con datos clasificados a priori, sino que basados en las características de los datos se busca agrupar aquellos que son similares. En el contexto de las transacciones PSE de Bancolombia, dado que a un conjunto de datos se le asignó una categoría, el problema se abordará mediante técnicas de clasificación supervisada.

En los problemas de clasificación supervisada mediante un conjunto de datos clasificados a priori, se reconocen patrones y características. Como objetivo, se busca que el modelo prediga con error pequeño nuevos datos, es decir, que pueda generalizar el conocimiento adquirido con los datos de entrenamiento en datos sin clasificar. Para esto, se han desarrollado diferentes técnicas, entre ellas regresiones lineales y logísticas, análisis discriminante, modelos bayesianos, Support Vector machines, árboles de decisión y redes neuronales.

Los modelos lineales construyen un predictor lineal para obtener probabilidades de clasificación binaria y se extienden varios clasificadores binarios para construir uno con clasificación múltiple. Por otro lado, las regresiones logísticas utilizan funciones logit para la predicción de la probabilidad de pertenecer a una de dos categorías. Esta técnica se puede ampliar para crear regresión multinomial logística, cuya intención es predecir las probabilidades de múltiples categorías a clasificar. En el caso de los clasificadores de Bayes, estos permiten conocer una probabilidad desconocida de clasificación basados en probabilidades conocidas sobres los datos.

Dentro de los clasificadores más complejos se encuentran Support Vector Machines y redes neuronales multicapa multinomiales. En la primera técnica la intuición es la representación de los

datos como puntos en un espacio dimensional para reconocer agrupamientos. Para esto es sumamente importante encontrar una función (Kernel) que proyecte adecuadamente los datos en el espacio construido. Por su parte, las redes neuronales son un sistema de neuronas interconectadas en una red que colabora para producir un estímulo de salida. Las redes construyen representaciones internas de los datos de entrada, con el fin de transmitir información valiosa para las siguientes capas. Internamente, las redes neuronales pueden utilizar diferentes funciones de modelación, por ejemplo, activación sigmoidal, tangente hiperbólica, función de base radial, entre otras.

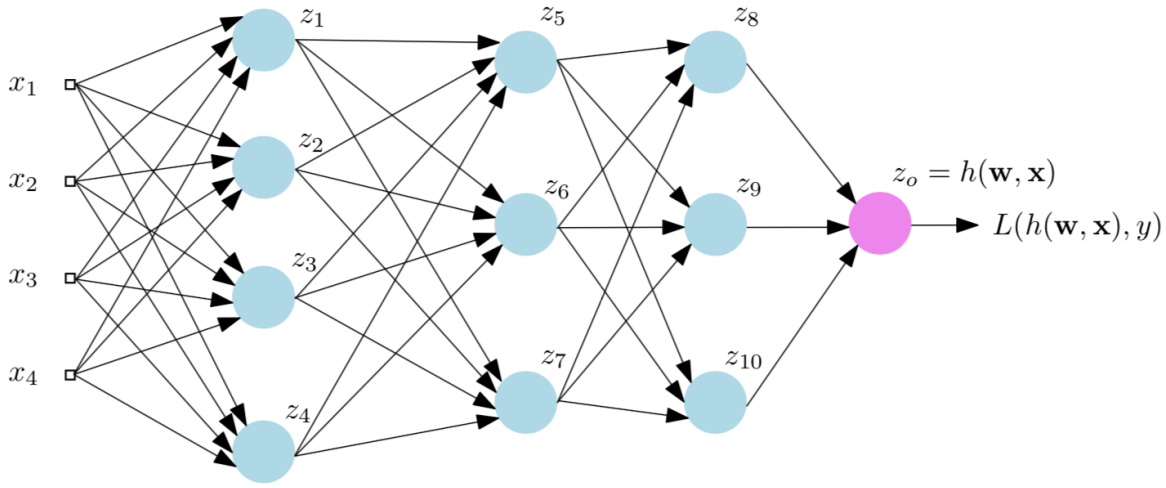

Figura 3. Modelo general red neuronal (Lozano, 2016).

En el modelo de redes neuronales existe un compromiso entre la complejidad de las funciones que se pueden implementar y el ajuste a los datos de entrenamiento. Una red muy simple puede no ser suficiente para capturar la estructura en los datos; no obstante, una red que modele funciones muy complejas puede sobre ajustarse a los datos y resultar en pobre generalización. Adicionalmente, el problema con redes muy complejas es que pueden existir muchas redes que se ajusten bien a los datos, lo cual implica que el modelo que se obtiene puede tener alta varianza.

Considerado las ventajas y desventajas de los modelos anteriores, tanto teóricas como de implementación, se utilizaron los modelos de regresión logística multinomial y redes neuronales multinomiales. El primero, por su capacidad de relacionar variables de entrada y probabilidades para asignar valores convertidos a categorías; y el segundo, por la capacidad de construcción de modelos basados en variables diversas como rangos de tiempo, valores numéricos y nominales.

6.1. Definición del Modelo

En primer lugar, se agregó la información de transacciones y clientes en una base de datos adicional. Luego, se normalizó la variable Valor transacción dividiéndola por el máximo y se crearon variables dummy de todas las variables categóricas. Esta base de datos se dividió en dos: registros con categoría de transacción y registros sin categoría de transacción. La primera base se utiliza para el entrenamiento y evaluación del modelo de clasificación. Posteriormente, el modelo predice la

categoría de transacción de la segunda base de datos. En la Tabla 3 se presenta la definición de los parámetros del modelo.

	Variable	Tipo
Predictores	Valor transacción	Numérica
	Dia semana	Dummy
	Mes	Dummy
	Fin de semana	Dummy
	Quincena	Dummy
	Momento del día	Dummy
	Segmento	Dummy
	Ocupación	Dummy
	Tipo vivienda	Dummy
	Nivel académico	Dummy
	Estado civil	Dummy
	Género	Dummy
	Edad	Dummy
	Ingreso rango	Dummy
Respuesta	Categoría transacción	Dummy

Tabla 3. Definición del modelo de clasificación

6.2. Modelo logístico multinomial

El modelo logístico multinomial se implementó en R. Se corrió en un equipo Core i7 de 8 núcleos y memoria RAM de 32 GB. La muestra de registros con categoría de transacciones se dividió aleatoriamente en dos, 60% para entrenamiento (1’840.918) y 40% para evaluación (1’227.276). Para medir la exactitud del modelo se calcularon el número de transacciones clasificadas correctamente:

	Muestra entrenamiento	Muestra evaluación
Exactitud (%)	30.13%	30.07%

Tabla 4. Porcentaje de exactitud del modelo multinomial logístico

6.3. Redes Neuronales

El modelo de redes neuronales se implementó utilizando el paquete Keras en Python 3.6. Se corrió en un equipo Core i7 de 8 núcleos y memoria RAM de 32 GB. La muestra de registros con categoría de transacciones se dividió aleatoriamente en dos, 80% para entrenamiento (2’454.555) y 20% para evaluación (613.639).

En la literatura se encontró que no hay un método óptimo para definir los nodos y capas de la red y sugieren plantear diferentes escenarios para calibrar estos escenarios; por lo tanto, se probaron 13 configuraciones y se calculó el porcentaje de registros clasificados correctamente en cada una (Tabla 4).

Configuración de la red*	Exactitud (%)	
	Muestra entrenamiento	Muestra evaluación
99/10/10/10/17	41.34%	41.32%
99/20/10/10/17	41.81%	41.84%
99/50/10/10/17	41.89%	41.96%
99/50/50/10/17	42.43%	42.38%
99/50/50/50/17	42.73%	42.75%
99/100/10/10/17	41.48%	41.45%
99/100/50/10/17	42.54%	42.48%
99/100/50/50/17	43.75%	43.75%
99/100/100/10/17	42.62%	42.61%
99/100/100/50/17	43.14%	43.13%
99/100/100/100/17	43.87%	43.87%
99/500/100/100/17	45.19%	45.06%
99/500/500/500/17	44.85%	44.70%

*Inputs/Nodos layer 1/ Nodos layer 2/Nodos layer 3/Número de categorías

Tabla 5. Porcentaje de exactitud en muestras de evaluación y entrenamiento por configuración de la red neuronal

Teniendo en cuenta lo anterior, se concluye que los mejores resultados de clasificación se obtienen con la configuración 99/500/100/100/17. Igualmente, se observa que el desempeño de las redes neuronales fue considerablemente superior al del modelo multinomial logístico. Es importante mencionar que también se analizó el porcentaje de exactitud por categoría y nuevamente los mejores resultados se obtuvieron con dicha configuración.

6.4. Resultados de la clasificación

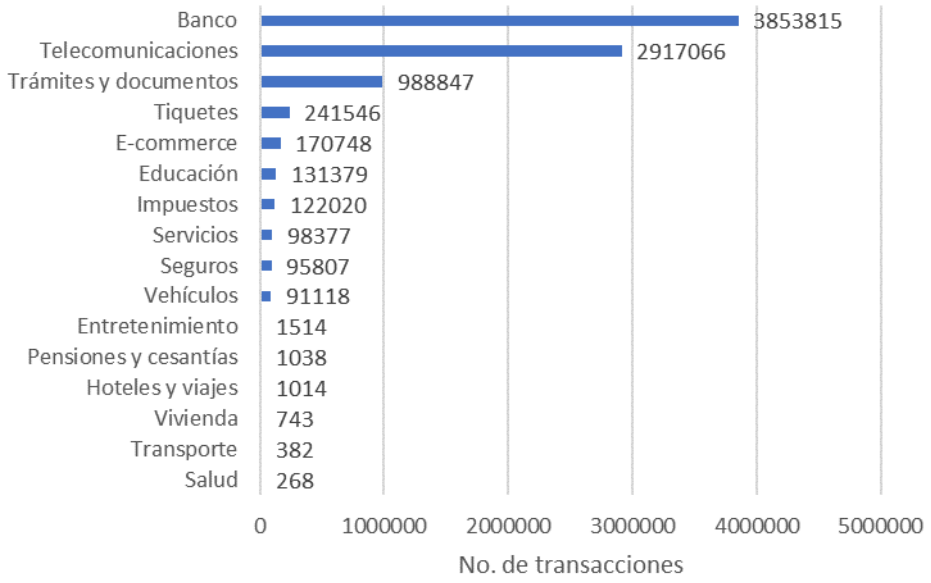

Figura 4. Resultados clasificación en transacciones no categorizadas

Una vez seleccionado el modelo de red neuronal 99/500/100/100/17 como el mejor clasificador, este se usó para predecir la muestra de transacciones sin categoría con 8 millones de registros (Figura 4). En este caso, no es posible determinar qué tan acertada es la clasificación.

7. Aplicaciones de la categorización

Como se mencionó previamente, la categorización de las transacciones puede permitir añadir funcionalidades a los PFMs. Por ejemplo, se puede predecir el gasto de un usuario discriminado para cada una de las categorías, de tal forma que este pueda decidir si debe tomar alguna medida sobre sus hábitos de consumo y en qué tipo de bienes y servicios debería hacerse. Para ilustrar cómo se podría llevar a cabo la implementación de esta función, planteamos un modelo de predicción del gasto anual en telecomunicaciones de los clientes.

Primero, calculamos el valor total de transacciones de telecomunicaciones en el 2017 por cliente, utilizando la base de 3 millones de registros categorizados. Para esto, se tuvo en cuenta únicamente los clientes con transacciones de telecomunicaciones mayores a cero. Por otro lado, como predictores definimos las variables demográficas de los clientes con el mismo tratamiento de limpieza y categorización que se realizó para el modelo de clasificación. En total se consolidó la información de 55.785 clientes.

Luego, evaluamos un modelo de regresión lineal múltiple y un modelo de Random Forest. Este último es una técnica basada en árboles, en donde el espacio de los predictores se divide en diferentes regiones. En Random Forest, se crean múltiples árboles y se promedia el resultado de cada uno para una predicción final. Para el entrenamiento de los modelos se utilizó una muestra con el 80% de los clientes, seleccionados aleatoriamente; el 20% restante se utilizó para la evaluación.

Como resultado, la raíz del error cuadrático medio en la muestra de test fue de 423.581 en el modelo de Random Forest, y 420.558 en el modelo de regresión lineal múltiple. En este caso tuvo un mejor resultado el modelo lineal; sin embargo, existe una amplia variedad de modelos de predicción que no se pudieron evaluar por las limitaciones del tiempo.

8. Conclusiones

Sobre la experiencia que se obtuvo al seguir la metodología se puede concluir lo siguiente:

- Las primeras etapas brindaron una visión más clara del significado de los datos y permitió establecer métodos para asegurar que la información fuese valiosa.
- El análisis lexicográfico permitió agrupar un conjunto de palabras que describen una parte considerable de los datos. Esto permitió generar una clasificación inicial de transacciones, que serían útiles para la construcción de modelos de clasificación.
- Se requirió de un conocimiento local sobre las palabras para realizar su clasificación, debido a la presencia de nombres de empresas y variaciones propias del lenguaje.

- Con el objetivo de encontrar un modelo que clasificará todas las transacciones en las categorías elegidas, se agregó la información de los clientes a las transacciones. Este proceso fue riguroso y requirió validaciones de que el procedimiento fue correcto.
- Se probaron dos modelos de la clasificación: logístico multinomial y redes neuronales con 13 diferentes configuraciones. Los mejores resultados de clasificación se obtuvieron con el modelo de redes neuronales, el cual recomendamos para trabajos futuros.
- La categorización de las transacciones permite agrupar, analizar y procesar la información para construir valor a los clientes y al banco. Un ejemplo es la posibilidad de predecir el gasto de los clientes en cada una de las categorías, el cual ilustramos mediante la predicción de los gastos anuales en telecomunicaciones.

9. Bibliografía

Corso, C. (s.f.). Aplicación de algoritmos de clasificación supervisada usando Weka. Universidad Tecnológica Nacional, Facultad Regional Córdoba. Recuperado el 26 de octubre 2018 de http://www.investigacion.frc.utn.edu.ar/labsis/Publicaciones/congresos_labsis/cynthia/CNIT_2009_Aplicacion_Algoritmos_Weka.pdf

James, G. et al. (2013). An Introduction to Statisitical Learning with Applications in R. Springer.

Lozano, F. (2016). Notas de clase Fernando Lozano. Universidad de los Andes.

Revista Dinero. (2 de agosto de 2018). Más de 188.676 transacciones electrónicas se realizaron en Colombia. Recuperado el 26 de octubre de 2018 de https://www.dinero.com/economia/articulo/transacciones-electronicas-realizadas-en-colombia/255163