Clasificación de transacciones ejecutadas a través de PSE

Esteban Betancur Valencia
Miguel Angel Mejía
Sebastian Rodriguez Colinas
29 de octubre de 2018

1. Introducción

El presente informe detalla el proceso de implementación de una metodología para clasificar transacciones de clientes de Bancolombia a través de la plataforma PSE del ACH. Este proyecto está alineado con la intención de mejorar las capacidades de analítica de datos del Banco y en especial ofrecer a los clientes información útil a partir de los datos recopilados.

El informe parte de la descripción de los datos, donde se listan las variables disponibles y se analizan las variables principales. Posteriormente en la metodología se definen las categorías escogidas y se propone una clasificación en dos etapas: se definen reglas de negocio para asignar categorías a una porción de los datos y luego se implementa un modelo de Random Forest para clasificar el resto de las transacciones entrenándolo con los datos ya etiquetados. Finalmente se exponen los resultados principales.

2. DATOS

Los datos provistos están organizados en dos tablas vinculadas, la primera (archivo dt_-trxpse_personas_2016_2018_muestra_adjt.csv) lista 11.866.544 transacciones a través de la plataforma PSE realizadas por personas naturales a empresas, entre las fechas 01-Sep-2016 y 01-Oct-2018. Esta lista contiene para cada transacción los siguientes atributos:

1. id_trn_ach: Número único de identificación de la transacción asignado por la plata forma PSE
2. id_cliente: Número que identifica el cliente que realizó la transacción
3. fecha: Fecha de la transacción
4. hora: Hora de la transacción
5. valor_trx: Valor en pesos de la transacción
6. ref1: Primera cadena de texto asignada por la empresa recaudadora a la transacción
7. ref2: Segunda cadena de texto asignada por la empresa recaudadora a la transacción
8. ref3: Tercera cadena de texto asignada por la empresa recaudadora a la transacción
9. sector: Sector económico al que pertenece la empresa recaudadora de la transacción, asignado por Bancolombia
10. subsector: Subsector económico al que pertenece la empresa recaudadora de la transacción, asignado por Bancolombia
11. descripción: Descripción del tipo de transacción realizada, asignada por Bancolombia

La segunda tabla corresponde a la información de cada uno de los clientes que realizó la transacción (archivo dt_info_pagadores_muestra.csv). Esta tabla lista diferentes atributos de 338.606 usuarios diferentes. Las columnas de esta tabla son:

1. id_cliente: Número que identifica el cliente, sirve de vínculo para la tabla de transacciones
2. seg_str: Segmento estratégico que le asigna el banco a este cliente
3. ocupación: Clasificación que le da Bancolombia a la ocupación del cliente
4. tipo_vivienda: Clasificación que le da Bancolombia al tipo de vivienda del cliente
5. nivel_academico: Clasificación que le da Bancolombia al nivel académico del cliente
6. estado_civil: Clasificación que le da Bancolombia al estado civil del cliente
7. genero: Masculino o Femenino
8. edad: Edad en años
9. ingreso_rango: : Clasificación que le da Bancolombia a sus ingresos de acuerdo a un rango

2.1. BASE DE DATOS EN FORMATO SQL

Como primer paso se creó una base de datos SQL en la plataforma MICROSOFT AZURE. se consolidaron las dos tablas y se crearon diferentes vistas para acceder a ellas desde cualquier plataforma. Para la carga de los datos se utilizó el wizard de importación de datos de Microsoft SQL Server Management Studio. La creación de la base de datos cosideró los siguientes pasos:

1. Creación de servidor (Hosting) SQL en MICROSOFT AZURE
2. Reemplazo de todos los caracteres "\n"seguidos por un caracter Alfa ([a-zA-Z]) por caracter vacío (NULL)

3. Carga de las tablas en formato *.csv
4. Creación de tablas nuevas donde se van a guardar los datos (con el tipo de dato correspondiente a cada columna)
5. INSERT de las tablas de datos originales a las tablas nuevas
6. Imputacion de edad a los usuarios que no tuvieran este atributo, se le asigna una edad sintética calculada como el promedio de la edad de los usuarios en ese mismo segmento y rango de ingresos
7. Creación de una vista consolidada, donde se asocia a cada usuario la información de su transacción correspondiente

Por seguridad y confidencialidad de los datos se le asignó al Host de la base de datos una restricción de direcciones IP y un login con contraseña para el acceso. Despues de la limpieza de los datos, resultaron 11.853.782 transacciones, es decir 12.762 transacciones menos que los datos originales. Estas transacciones removidas corresponden al 0,1 % del total. Con respecto a los usuarios resultaron 338.606, el mismo número que la tabla original.

2.2. Análisis de los datos

Una vez organizados los datos en la base de datos, se llevó a cabo un análisis preliminar de ellos. De acuerdo a la necesidad principal de este proyecto, se buscó una descripción de las variables principales que pueden ser usadas para la clasificación de las transacciónes.

Se analizó la correlación entre de cada uno de los dígitos del identificador único de las transacciones (id_trn_ach) y el subsector esperando que tuviera una cadena de números asociada a la empresa recaudadora. No se encontró ninguna correlación.

A continuación se describen las generalidades de las variables principales identificadas.

2.2.1. Sector, subsector y descripción

Estas tres variables asignadas a las transacciones son impuestas por el Banco de acuerdo a la caracterización que tiene de algunas empresas recaudadoras. Las empresas deben ser clientes del banco para conocer su información. Del total de los datos, solo 3.307.385 tienen estos atributos, correspondientes al 28 % como se ilustra en la Figura 2.1.

La variable 'Sector' tiene 10 categorías, 'Subsector' 54 categorías y 'Descripción' 160. Estas variables se tomarán como la primera opción para asignar una clasificación a las transacciones.

2.2.2. Ref1, ref2 y ref3

Estas son variables que la empresa recaudadora asigna como referencia de la transacción, en general son texto libre pero contienen palabras clave que pueden ser útiles para la clasificación. 11.492.514 transacciones (97 % de los datos) tienen asignado algún texto en la variable ref1, 6.790.150 (57 %) tienen asignado un texto en ref2 y ninguna transacción tiene texto en la variable ref3.

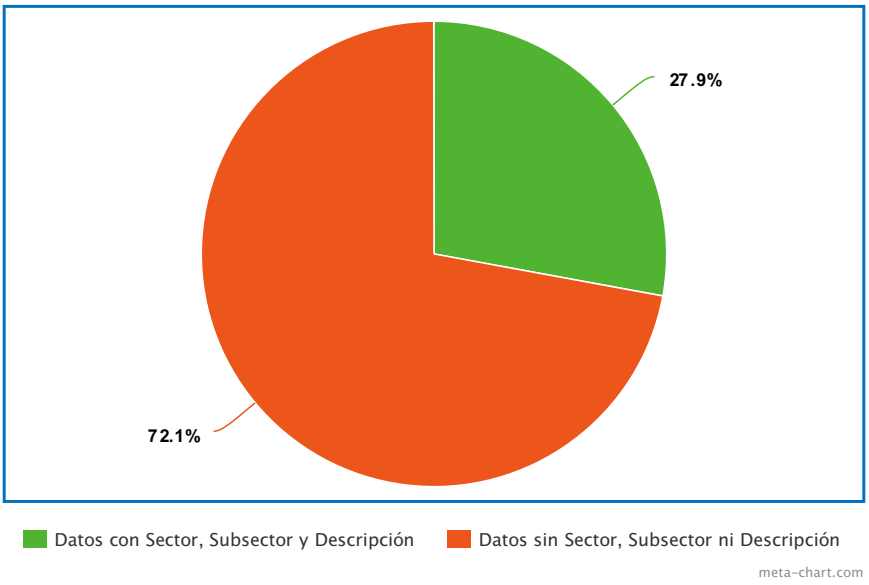

Figura 2.1: Distribución de los datos con Sector, Subsector y Descripción

3. METODOLOGÍA

La metodología se divide en dos etapas principales: definición de las categorías y categorización de las transacciones.

3.1. DEFINICIÓN DE LAS CATEGORÍAS

Las diferentes categorías de clasificación resultan de una combinación de diferentes propuestas. La primera propuesta de clasificación es la entregada por el Banco:

1. Comida
2. Hogar
3. Cuidado personal
4. Entretenimiento
5. Educación
6. Transporte
7. Viajes
8. Ahorro
9. Pago de deudas
10. Ingresos
11. Retiros en efectivo
12. Mascotas
13. moda

14. Tecnología y comunicaciones
15. Otros

Por otro lado, se analizan las clasificaciones de las empresas recaudadoras que tiene la plataforma PSE:

1. Portal de Pagos Electrónicos - Banco de Bogotá
2. Multipagos PSE – Bancolombia
3. Centro de Pago PSE - Banco de Occidente
4. Comercio
5. Educación
6. Entretenimiento
7. Financiero
8. Gobierno
9. Otros Servicios
10. Salud
11. Transporte
12. Vivienda
13. Tecnología y Comunicaciones
14. Servicios Públicos y TV Por Cable

De acuerdo al análisis de las dos listas de categorías, se define una lista combinada de 16 items que resulta principalmente de remover las categorías 'Ingresos', 'retiros en efectivo' y 'otros' de la lista propuesta por Bancolombia y agregar 'seguros', 'salud', 'gobierno' y 'servicios públicos y tv por cable'. A continuación se listan las categorías finales escogidas con algunas descripciones para evitar ambigüedades:

1. Comida
2. Vivienda: equivalente a 'Hogar', incluye arriendos, pagos a constructores, administración, entre otros
3. Cuidado personal
4. Entretenimiento
5. Educación: Incluye el pago de colegios que comúnmente es un gasto asignado a los hijos
6. Transporte: atribuido principalmente a los vehículos, no incluye el transporte escolar, ni vuelos
7. Viajes: incluye vuelos y hoteles principalmente
8. Ahorro: Pagos a pensión y otros ahorros

9. Pago de deudas: Pagos de tarjetas de crédito
10. Mascotas
11. Moda: Todo lo relacionado con prendas de vestir
12. Tecnología y comunicaciones: Incluye compra de equipos y pago de mensualidades
13. Seguros: Pago de polizas sin incluir las de salud ni vehiculares
14. Salud: Gastos correspondientes a salud y pago de polizas de salud
15. Gobierno: Pagos de impuestos e instrumentos públicos
16. servicios públicos y tv por cable

3.2. Clasificación de las transacciones

Para clasificar las transacciones en diferentes categorías se propone la siguiente metodología:

1. Asignar categorías mediante la definición de reglas de negocio justificadas
2. Proponer una estructura de machine learning para clasificar cualquier transacción que no cumpla las reglas de negocio del paso 1

A continuación se explican las dos metodologías utlizadas.

3.2.1. Reglas de negocio

A partir del análisis de los datos se definen como variables principales para crear reglas de negocio el 'Subsector' y la 'ref1'. Aunque parezca un método rudimentario", se imputarán categorías con una justificación válida para confiar en las etiquetas asignadas.

De acuerdo al análisis de los atributos presentes en 'Subsector', se encuentra que algunos de los subsectores pertenecen por completo a alguna de las categorías por la homogeniedad de las transacciones asignadas. Otros subsectores en cambio no pueden ser definidos ya que incluyen transacciones con más variación. La lista de subsectores con categorías asignadas y su justificación se presenta en la tabla 3.1:

Se aprecia que se logra clasificar con esta regla de negocio 2.124.851 transacciones (18% del total). Aunque es un buen punto de partida para tener datos etiquetados, se nota además que solo se tienen etiquetas de 9 de las 16 categorías totales. No tener etiquetas de todas las categorías llevaría a problemas con cualquier método de clasificación usando ML.

La segunda variable utilizada para definir reglas de negocio es la 'ref1'. Aunque esta variable es muy heterogénea en los datos, se evidencian palabras clave que pueden definirla categoría de las transacciones. Usando herramientas de análisis de texto, se obtienen las 100 palabras más frecuentes en toda esta columna descartando las más comunes que no entregan información (como pago, cc, compra, psepayment, entre otras). De esta lista de palabras, se escogen

Subsector	Num Records	Etiqueta Impuesta	Justificacion
TRANSPORTE AEREO	113	Viajes	vuelos
HOTELES	943	Viajes	Gasto propio de viajes
CEMENTO	2723	Vivienda	Corresponde a pagos a constructores
FERRETERÍA	243	Vivienda	Corresponde a compras para el hogar
VALOR AGREGADO	548028	Tecnología y comunicaciones	Todas corresponden a Actividades de telecomunicaciones inalÃ¡mbricas
ESTABLECIMIENTOS EDUCATIVOS	39611	Educacion	
MUNICIPIOS	95742	Gobierno	
ADMINISTRACIÓN CENTRAL	331832	Gobierno	
AUTOMOTORES	148	Transporte	
OBRAS DE INFRAESTRUCTURA	18572	Transporte	Todas son por recaudo de peaje
COMBUSTIBLES Y LUBRICANTES	419	Transporte	Gasto asignado a transporte
COMERCIO DE REPUESTOS	17	Transporte	
EPS Y SALUD PREPAGADA (SALUD)	11759	salud	Empresas de planes de salud
ACUEDUCTO Y ALCANTARILLADO	84	servicios publicos y tv por cable	
TELEFONIA FIJA	623771	servicios publicos y tv por cable	Telefonia fija se asigna a servicios publicos
ELECTRICIDAD	442489	servicios publicos y tv por cable	Servicio publico
SEGUROS	8357	Seguros	
Total	**2124851**		

Cuadro 3.1: Subsectores utilizados para asignar categorías

las más pertinentes para asignar una categoría, entonces si alguna de las palabra escogidas está presente en la 'ref1', se asignará la la categoría definida. La lista de palabras junto con su categoría y justificación se presenta en la tabla 3.2:

Finalmente se asignan las categorías de acuerdo a estas dos dos reglas definidas.

3.2.2. ALGORITMO DE CLASIFICACIÓN DE TRANSACCIONES

Para clasificar las transacciones que no cumplen con las reglas de negocio definidas anteriormente y cualquier transacción en general, se propone utilizar un algoritmo de Machine Learning llamado Random Forest, se escoge este algoritmo por sus buenos resultados en diferentes competiciones de clasificación (por ejemplo en https://www.kaggle.com/competitions), además por su capacidad de recibir variables categóricas y numéricas. Aunque el overfitting es una inconveniente de este método, la gran cantidad de datos puede mantenernos lejos de caer en ello.

Como los datos de entrenamiento de este algoritmo vendrán generados de acuerdo a las reglas, las variales ref1 y subsector no se usarán como datos de entrada del modelo. Para más precisión del modelo sería posible usar ref1 quitando de todas las filas las palabras que se usaron para la regla, pero las limitaciones de tiempo del proyecto impidieron hacer esta implementación.

Las variables de entrada (X) del modelo para clasificar las transacciones son:

1. seg_str: Segmento del cliente que realizó la transacción
2. ocupacion: Ocupación del cliente que realizó la transacción
3. tipo_vivienda: Tipo de vivienda del cliente que realizó la transacción
4. estado_civil: Estado civil del cliente que realizó la transacción
5. genero: Genero del cliente que realizó la transacción
6. edad: Edad del cliente que realizó la transacción
7. ingreso_rango: Rango de ingresos del cliente que realizó la transacción

Palabra en ref1	Categoria asignada	Justificacion
avianca	vajes	Empresa que vende vuelos
administracion	vivienda	normalmente de conjuntos residenciales
administraci	vivienda	normalmente de conjuntos residenciales
celular	Tecnología y comunicaciones	Pago de mensualidad o compra de celulares
colegio	Educacion	Pago de mesualidad de educacion
apto	vivienda	normalmente admon de conjuntos residenciales
flight	viajes	Vuelo
impuesto	Gobierno	Pago de impuestos al gobierno
impuestos	Gobierno	Pago de impuestos al gobierno
curso	Educacion	Pago de algun tipo de educacion
departamento	Gobierno	Pago de impuestos al gobierno departamental
diplomado	Educacion	Pago de un diplomado
pension	ahorro	Se toma pension como un ahorro
pensi	ahorro	Se toma pension como un ahorro
soat	Transporte	Compra del soat de un vehiculo
vuelos	viajes	Vuelo
arrendamiento	vivienda	Arriendo de vivienda
avantel	Tecnología y comunicaciones	Empresa de celulares
camiseta	Moda	Articulo de vestir
desayuno	Comida	Palabra que pertenece a alimentación
grado	Educación	Palabra relacionada con educación
noches	viajes	Usualmente pago de un hotel
ropa	moda	Palabra correspondiente a artículos de vestir
tiquetesbaratoscom	viajes	Empresa que vende vuelos
transporte	Transporte	Palabra igual a la categoría
tarjeta de credito	Pago de deudas	Pago de cuota de tarjeta de credito
salud	Salud	Palabra igual a la categoría
gafas	moda	Articulo de vestir
pantalon	moda	Articulo de vestir
crema	cuidado personal	Palabra correspodiente a artículos de cuidado personal
tratamiento	cuidado personal	Palabra correspodiente a artículos de cuidado personal
hotel	hotel	Palabra correspodiente a viajes
gas	servicios publicos y tv por cable	Palabra correspodiente a servicios publicos
emcali	servicios publicos y tv por cable	Empresa que vende sevicios publico
predial	Gobierno	Pago de impuesto predial
cine	entretenimiento	Actividad propia de entrenetimiento
ciudaddemascotascom	mascotas	Empresa que vende productos para mascotas

Cuadro 3.2: Palabras presentes en ref1 utilizadas para asignar categorías

8. fecha: Fecha de la transacción

9. valor_trx: Valor de la transacción

10. promedio_transaccion_usuario: Promedio de las transacciones pasadas a través de PSE del cliente que realizó la transacción

11. promedio_anual_transacciones: Promedio anual de transacciones pasadas a través de PSE del cliente que realizó la transacción

12. num_transacciones: Número de transacciones pasadas a través de PSE del cliente que realizó la transacción

13. num_anual_transacciones: Número de transacciones anuales pasadas a través de PSE del cliente que realizó la transacción

14. total_transacciones: Monto total transado a través de PSE del cliente en el momento que realizó la transacción

Se aprecia que se proponen nuevas variables que dependen de la información de transacciones pasadas del mismo cliente, aunque algunas pueden ser mas relevantes que otras para el modelo, el mismo entrenamiento define la importancia de cada una.

La salida del modelo será la categoría más probable a la que pueda pertenecer una transacción.

El preprocesamiento de los datos incluye la conversión de las variables categóricas ('seg_str', 'ocupacion', 'tipo_vivienda', 'estado_civil', 'genero']) a matrices binarias y la separación de los datos en entrenamiento y prueba. La selección de hiperparámetros de midelo se realiza a través de un Grid Search con cross validation.

4. RESULTADOS

La implementación del método no se logró hasta el momento. Principalmente el tamaño de los dato aumentó los tiempos de ejecución de cada algoritmo, llevándonos al deadline sin un resultado final.

No obstante, la metodología presentada y los archivos de python entregados muestran el gran avance del proyecto y la metodología planteada.