DATATÓN BC 2018

Hernan David Carvajal
Carlos Antonio Pinzón
Miguel Ángel Romero

Octubre 2018

1. Descripción del problema

Los clientes de Bancolombia realizan transacciones en linea a través del servicio PSE. Este servicio permite a las empresas ofrecer a sus clientes la posibilidad de realizar pagos y/o compras, debitando los recursos en línea de la Entidad Financiera donde el cliente tiene su dinero y depositándolos en la Entidad Financiera recaudadora que defina la Empresa o Comercio [1].

Las transacciones que se realizan en PSE no pueden ser clasificadas según el sector o la naturaleza de la misma debido a que no se cuenta con información precisa del pago o compra que realiza el cliente. A diferencia de las transacciones realizadas en PSE, las transacciones realizadas en puntos de venta (POS) son clasificadas de acuerdo al estándar ISO 18245, el cual esta relacionado con el tipo de bien o servicio que la empresa provee.

Clasificar los gastos de un cliente es importante para la gestión de las finanzas personales y la promoción de productos y servicios financieros por parte de Bancolombia.

1.1. Objetivo

Clasificar las transacciones de clientes de Bancolombia realizadas a través del servicio virtual PSE entre Agosto del 2016 y Agosto del 2018.

2. Resumen de la solución

Definamos los siguientes tipos de transacciones: (I) aquellas que contienen texto suficiente como para ser categorizadas; y (II) aquellas que no.

- Para las de tipo I, desarrollamos un método explícito de clasificación por palabras clave en el que se estima mediante un pequeño sistema de votación sobre el texto de la transacción, a qué categoría pertenece.
- Para las de tipo II utilizamos un modelo random forest entrenado y validado sobre el conjunto de todas las transacciones de tipo I (que ya ha sido etiquetado). El soporte teórico detrás de la idea de utilizar las tipo I para entrenar el modelo de las tipo II, es que toda transacción de tipo I se vuelve tipo II cuando se ignora el texto que la acompaña.

3. Solución del problema

La solución del problema consta de tres partes: la primer parte corresponde a la limpieza general de los datos de clientes y transacciones; en la segunda parte se realizó la clasificación por palabra clave de las transacciones que tuvieran suficiente información en los campos de texto; y en la última parte se desarrolló el modelo que permite la clasificación de las transacciones que no contengan texto suficiente.

3.1. Limpieza general de los datos

3.1.1. Conjunto de transacciones

En el set de transacciones encontramos líneas en las que aparecían más comas de las que debían, en su mayoría porque hacían parte de la descripción, e.g. Industria de metales, vidrios y materiales de construcción. También hubo líneas en las que aparecían comillas dobles que no cerraban nunca.

Todas las líneas en las que ocurría alguna de las dos problemáticas fueron removidas del conjunto. Éstas conformaban apenas un 0.40 % del total.

Posteriormente,

1. tomamos todas las palabras y frases incluidas en los campos ref1, ref2, ref3, sector, subsector y descripción, y las juntamos en una sola cadena;
2. reemplazamos todos los caracteres no alfabéticos por espacios;
3. reemplazamos todas las tildes y eñes por su caracter equivalente a-z o A-Z (esto lo hicimos porque para homogeneizar las palabras, es más fácil quitar a las tildes a todas que ponerlas donde corresponde);
4. convertimos todas las letras a minúsculas;
5. convertimos las palabras en plural a singular;

6. juntamos todas las palabras utilizando un único espacio de separación entre cada par;

7. sobre el conjunto de datos, reemplazamos las columnas ref1, ref2, ref3, sector, subsector y descripción, por una única llamada “descripción”, que corresponde a la cadena limpia del paso 5.

Cabe decir que muchas transacciones quedaron con una descripción muy corta, como “cc”, “cc cc”, “nit”, “ce”, o “” (vacía).

3.1.2. Conjunto de clientes

En el set de clientes encontramos que algunas columnas contenían valores no incluidos dentro de la documentación, valores redundantes e inconsistencias con la edad.

Los datos categóricos que no hacían parte de las categorías preestablecidas fueron reemplazados por el valor nulo. Esto sucedió específicamente con el tipo de vivienda, en donde habían valores numéricos no incluidos en la descripción del problema. También se encontraron valores redundantes dentro de los datos, por ejemplo, en el nivel académico los valores ‘H’ y‘B’ representan lo mismo.

En cuanto a la edad, existen valores inconsistentes, los valores menores que 5 y mayores que 100 fueron reemplazados por nulo, puesto que consideramos que estas edades representan un outlier en el set.

3.2. Clasificación por palabras clave (cuando es posible)

En resumen hicimos lo siguiente.

- Definimos las categorías mostradas en la tabla 2
- Generamos un diccionario que contiene, para cada categoría, un listado de palabras clave asociadas a esa categoría. Una misma palabra puede estar asociada a más de una categoría.
- Usamos el diccionario para categorizar transacciones mediante un sistema de votación: si la mayoría de palabras del texto de la transacción están asociadas a una misma categoría, ésta se selecciona.

3.2.1. Funcionamiento del diccionario

El diccionario contiene, para cada categoría, un listado de palabras clave asociadas a esa categoría, y permite que hayan palabras relacionadas con más de una categoría. Su formato es el siguiente.

```
{
    categoría 1: lista de palabras relacionadas con la categoría 1,
    categoría 2: lista de palabras relacionadas con la categoría 2,
   ...
}
```

El método para clasificar una transacción T es el siguiente:

1. por cada categoría k, se cuenta la cantidad $f_k(T)$ de palabras que aparecen en T y que están asociadas a la categoría k;
2. luego se normaliza $f_k(T)$ en un nuevo valor $\hat{f}_k(T)$ para evitar que las categorías que en promedio tienen altos valores de $f_k(T)$ queden siempre escogidas;
3. y finalmente se calcula la categoría de T como $c(T) := \arg\max_k \hat{f}_k(T)$, salvo si $\max_k \hat{f}_k(T) = 0$, en cuyo caso se obliga a que $c(T) := -1$.

La normalización se hace por cada categoría con respecto al $f_k(T)$ promedio. Es decir, $\hat{f}_k(T) := \frac{f_k(T)}{\Sigma_t f_k(t)}$.

3.2.2. Método iterativo para la construcción del diccionario

El formato del diccionario es el siguiente.

```
{
    categoría 1: lista de palabras relacionadas con la categoría 1,
    categoría 2: lista de palabras relacionadas con la categoría 2,
   ...
}
```

En él pueden haber palabras relacionadas con más de una categoría.

El diccionario empezó con las categorías sugeridas, y para cada una, con al menos 3 palabras relacionadas.

Luego iniciamos el siguiente proceso iterativo para enriquecer el diccionario:

1. Categorice las transacciones utilizando el diccionario.
2. Reporte cuántas transacciones fueron categorizadas.
3. Reporte para cada categoría k, cuáles de las palabras contenidas en las transacciones categorizadas como k no hacen parte aún de la lista asociada a la categoría k.
4. Añada manualmente por cada categoría las palabras que estén realmente relacionadas y que valgan la pena.
5. Repita 1-5 hasta estar satisfecho.

Por ejemplo, supongamos que al iniciar, la única palabra asociada a la categoría VIAJES es hotel, y que la siguiente lista de transacciones hace parte del dataset:

Entonces ocurriría lo siguiente:

- En el paso 3, las palabras alojamiento (2) y pago (2) se reportan como posibles candidatas entre muchas otras que suelan acompañar a la palabra hotel. Los números entre paréntesis indican la cantidad de veces que aparece cada una acompañando a la palabra hotel.

Pago por alojamiento en el hotel Mi Tierrita
Hotel Decameron, pago alojamiento 2 días
Hospedaje Decameron Habitación 201
Pago habitación 402, alojamiento 3 días
Alojamiento 5 días Decameron
Pago arriendo habitación 905
Curso de artes 20 días

Cuadro 1: Ejemplo de categorización

- En el paso 4, se añade manualmente la palabra alojamiento (entre otras), pero no la palabra pago, porque sólo la primera está directamente relacionada con la categoría viajes.
- En el paso 5 se decide repetir porque se añadió una nueva palabra.
- En el paso 3 se reportan ahora pago (3), días (2), habitación (1), Decameron (1), entre otras.
- En el paso 4 se añaden habitación y Decameron.
- Así sucesivamente hasta completar la lista [hotel, alojamiento, habitación, Decameron, hospedaje].

3.2.3. Método iterativo para la selección de categorías

Encontramos los siguientes inconvenientes con las categorías propuestas:

- La categoría mascotas tenía muy pocas transacciones asociadas.
- La categoría ingresos, a diferencia del resto, no aplica para el comprador sino para el vendedor.
- Las transacciones relacionadas con salud, arrendamientos y pago de impuestos, entre otras, no corresponden a ninguna de las existentes (o corresponden a varias).

Por ello decidimos dividir, juntar y/o renombrar algunas categorías utilizando la siguiente metodología:

1. Categorice basado en las categorías iniciales.
2. De las transacciones que no fueron categorizadas, imprima las palabras frecuentes para ver qué categorías nuevas podrían considerarse.
3. Si hubo, añada las categorías relevantes encontradas en el paso 2 y repita 1-3.
4. Categorice basado en las nuevas categorías.

5. Para las categorías más pequeñas, revise semánticamente dentro cuáles se podría podría incluir.

6. Para las categorías más grandes, revise cómo se podrían dividir.

7. Si hay modificaciones pertinentes recomendadas por 5 y 6, hágalas y repita 4-7.

Al final resultaron las siguientes categorías:

Índice	Categoría
0	deudas
1	transferencias (a fondos de ahorro y de inversión, o para retiro en cajero (nequi))
2	educación
3	hogar
4	trámites (incluye expedición de certificados, pago de impuestos y de comparendos)
5	seguros
6	mercado moda
7	domésticos 1 (agua, gas y energía eléctrica)
8	domésticos 2 (telefonía fija, televisión e internet)
9	móvil (telefonía móvil: recargas y post-pago)
10	arriendo
11	viajes transporte
12	entretenimiento
13	otros
-1	(sin categoría)

Cuadro 2: Categorías de clasificación de transacciones

Y la gráfica de transacciones por categoría fue la siguiente:

3.3. Clasificación por modelo cuando no es posible por palabras clave

3.3.1. Entrenamiento del modelo: columnas añadidas

El set de datos de entrenamiento del modelo está compuesto por las transacciones que fueron clasificadas por palabras clave, y la unión del set de transacciones y clientes a través del identificador del cliente.

Encontramos necesario agregar nuevas columnas que brindaran mayor información acerca del historial de gastos de un cliente. Cuatro columnas fueron añadidas:

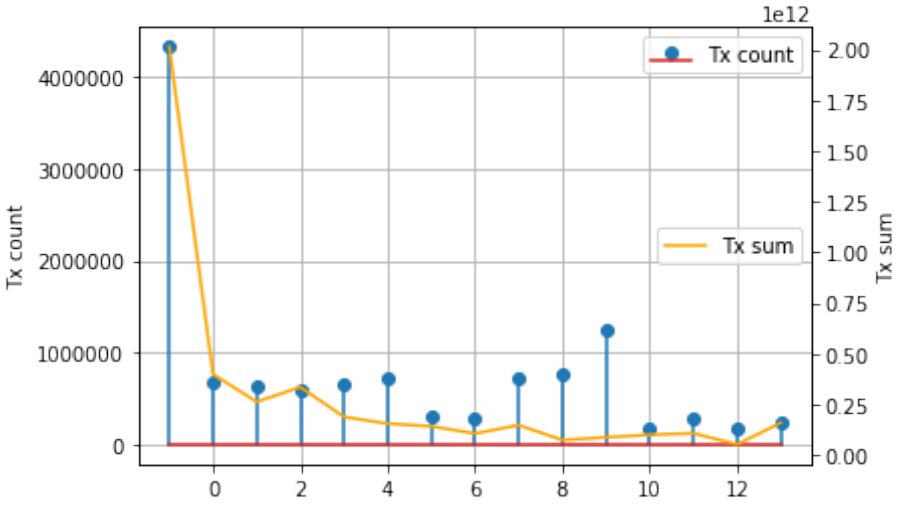

Figura 1: Cantidad (y valor total) de transacciones por categoría. -1 corresponde a transacciones sin categoría asignada.

- la primer columna corresponde a la categoría mas utilizada por el cliente entre las últimas (máximo 10) transacciones realizadas. Con esta columna buscamos identificar gastos consecutivos, tal como, gastos en moda o entretenimiento;
- la segunda columna corresponde a la categoría de la transacción mas cercana al mes anterior por el cliente, con el fin de identificar pagos periódicos, como servicios públicos o arrendamiento;
- las otras dos columnas añadidas corresponden al mes y el día la transacción que se está realizando, estas columnas son útiles para identificar la periodicidad de las transacciones.

Además de añadir 4 columnas al set de datos, también decidimos eliminar 3 columnas que generaban ruido en el modelo:

- la identificación de la transacción es un valor único que no esta relacionado con algún patrón en el set de datos,
- la identificación del cliente puede estar relacionado con un patrón, pero es muy difícil que el modelo pueda capturarlo, y además ya se añadieron columnas que relaciona el historial de transacciones del cliente,
- la fecha de la transacción fue reemplazada por el mes y día, puesto que no consideramos que el año brinde información útil al modelo.

Después de añadir y eliminar las columnas del set de datos, consideramos dos criterios para eliminar los valores nulos:

- eliminar las filas que contenían más de dos valores nulos,
- imputar los valores nulos restantes con la moda correspondiente a cada columna.

3.3.2. Desempeño: comparación con los primeros modelos

La técnica escogida para obtener el modelo fue "Random Forest Classifier". El primer modelo se obtuvo a partir del set de datos con las columnas originales y el criterio de imputación explicado anteriormente. Los resultados obtenidos de este modelo se pueden observar en la Figura 2

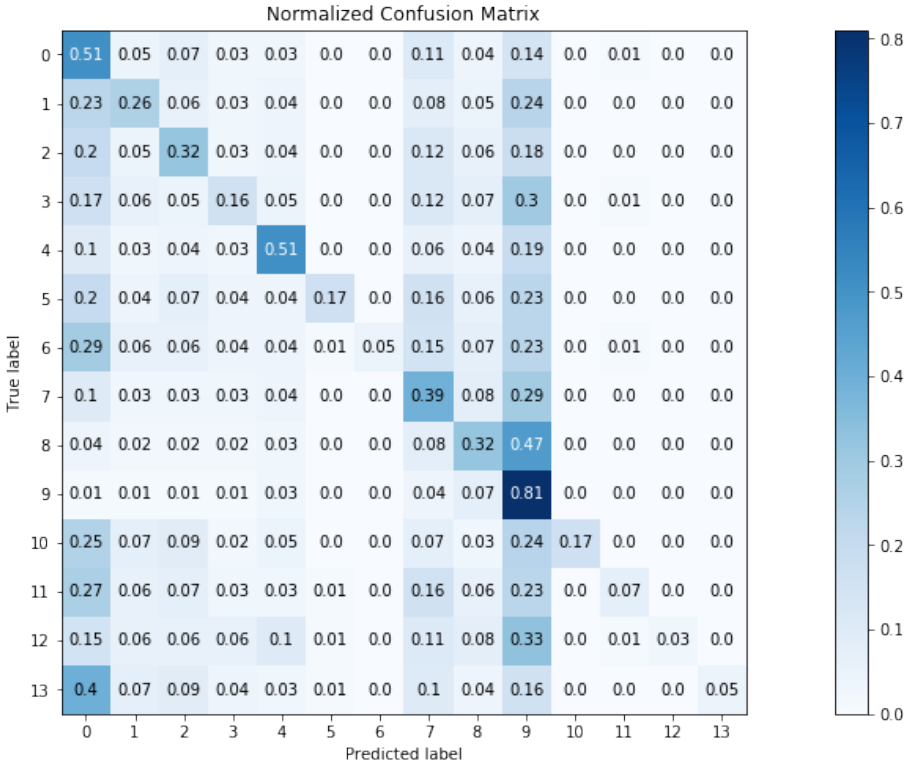

Figura 2: Matriz de confusión primer modelo

Las columnas más importantes en este modelo fueron 'id_trn_ach', 'id_cliente', 'fecha', 'hora', y 'valor_trx'. El identificador de la transacción es un valor único que no brinda información relevante para clasificar el tipo de transacción, pero el modelo le da gran importancia. Por esta razón, su resultado no es el esperado.

El modelo final se obtuvo a partir del set de datos con los cambios en las columnas, y la misma técnica con un total de 20 arboles (estimadores). Los resultados obtenidos de este modelo se pueden observar en la Figura 3.

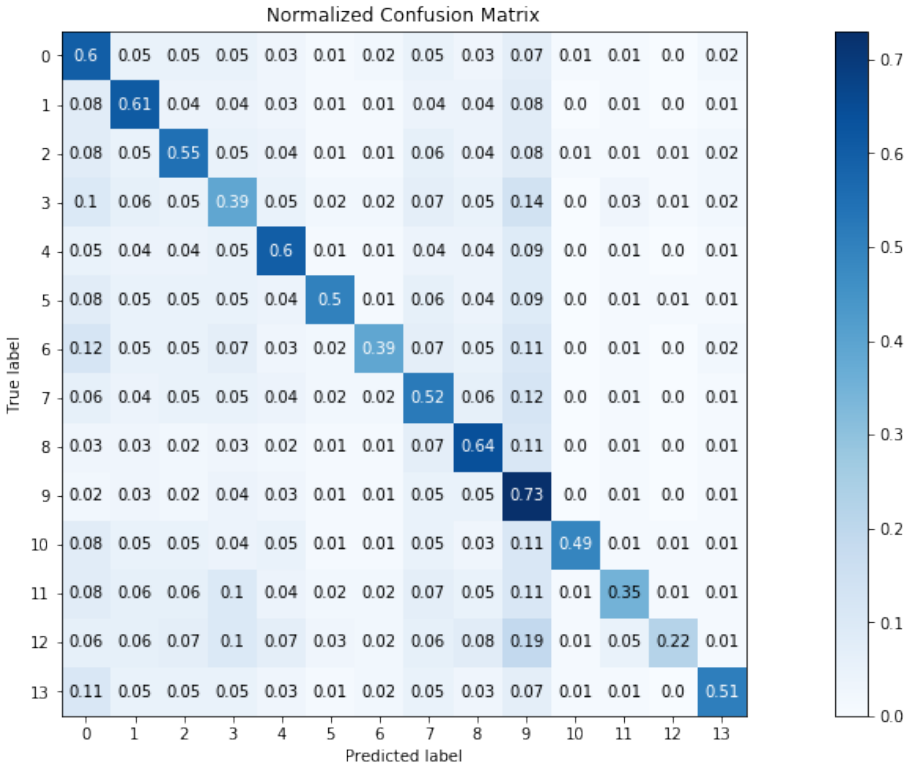

Figura 3: Matriz de confusión modelo final

Las columnas más importantes en este modelo fueron:

- 'valor_trx',
- 'hora',
- 'ingreso_rango',
- 'ultimas_trx',
- 'mes'.

Este resultado resalta la importancia de seleccionar las columnas adecuadas para clasificar las transacciones. En la sección final se listan algunas columnas que consideramos agregar al modelo.

3.3.3. Propuesta para implementación final

El modelo presentado anteriormente tenia en cuenta las transacciones previamente clasificadas iterativamente mediante el diccionario de palabras (aprendizaje supervisado). Dado que el objetivo del modelo es clasificar las transacciones que no fueron clasificadas mediante el diccionario, a continuación se describe el proceso que se debe realizar con estas nuevas transacciones para su clasificación y posterior inclusión en el set de datos:

1. Agregar la información de cliente y eliminar las columnas de identificación del cliente y de la transacción,
2. Agregar las columnas de mes y día, y eliminar la columna de fecha,
3. Agregar las columnas referente a la moda de las últimas 10 transacciones y del mes anterior,
4. Imputar los valores nulos con la moda de las columnas correspondientes,
5. Clasificar la transacción utilizando el modelo,
6. Agregar la transacción a set de datos con la clasificación obtenida del modelo.

4. Propuestas para trabajo futuro

Consideramos que el trabajo realizado puede extenderse y desarrollarse con más rigor para obtener mejores resultados. Mas específicamente, definimos tres tareas que demarcarían la exploración futura del proyecto en caso de ser seleccionado.

Tarea 1: Feature Engineering

Esta sería la primer tarea que desarrollaríamos.

En esta ocasión utilizamos (1) la moda de las últimas transacciones y (2) la transacción más cercana al punto temporal "exactamente hace un mes", sin embargo, otras posibles features candidatas son (3) la transacción moda para el usuario desde el inicio hasta hoy, (4) la transacción de precio más similar de entre las últimas hechas.

Tarea 2: Automatizar por completo la generación del diccionario y la selección de categorías utilizando técnicas de clustering sobre grafos

Se podría definir un grafo de palabras en el cual, dos palabras están relacionadas si y solo si existe una transacción en la que ambas aparezcan. A dicho arco podríamos asociarle un peso que represente la cantidad de transacciones en las que aparecen ambas.

Sobre el grafo eliminaríamos aquellos nodos u para los cuales $\Sigma_{(u,v,w)\in W}\ w$ supere algún umbral, pues estos nodos representan palabras genéricas como "a", "cc", "pago", "de", "con", "y", que aparecen en una gran porción de las transacciones.

Finalmente, exploraríamos técnicas que permitan encontrar automáticamente clusters de palabras relacionadas. Cada cluster correspondería a una categoría.

Tarea 3: Buscar clientes similares y explotar su similaridad

Si X es un cliente parecido a Y , y por un cierto periodo de tiempo se dificulta la tarea de clasificación de las transacciones de X , pero se conocen las de Y , se podría considerar la opción de utilizar la información de Y como información auxiliar adicional dentro de la clasificación.

Referencias

[1] PSE ¿como funciona? `https://www.pse.com.co/como-funciona`. Accessed: 2018-10-28.