Informe Datatón BC

Equipo: Los Predictivos

Introducción

Este informe tiene como propósito documentar los hallazgos a partir de los datos entregados por el banco en el marco de la competencia de analítica de datos “DatatónBC”. Para lograr lo anterior, se utiliza las aplicaciones de SAS: SAS Guide para la manipulación de datos, SAS Visual Analitycs (SAS Viya) para la visualización y análisis de los datos y SAS Enterprise Miner para la modelación de los datos.

Teniendo en cuenta lo anterior, este documento se divide en 3 partes. En primer lugar, se muestran los resultados de la exploración y limpieza de datos para luego mostrar los resultados de los modelos aplicados y finalmente se concluye.

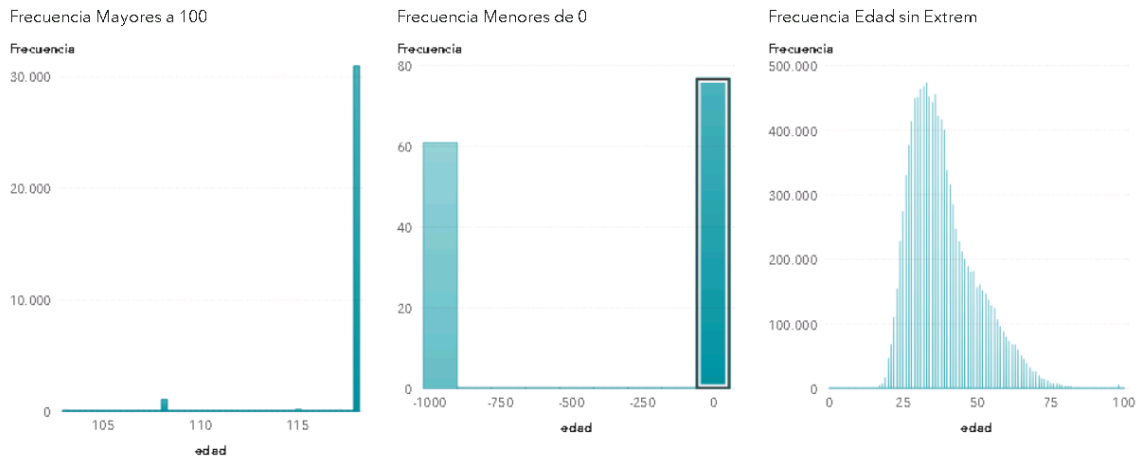
1. Exploración y limpieza

1.1. Limpieza de datos

El primer proceso realizado durante el proyecto fue unir las tablas de transacciones y pagadores utilizando SAS Enterprise Guide usando como campo llave la columna de id_cliente. Inmediatamente, se organizaron las fechas que tenían problemas, se crearon variables de mes, día, semana y año con el fin de analizar las series de tiempo de las variables de monto de transacción.

Se crearon variables adicionales, como: el ingreso máximo e ingreso mínimo, con el fin de ser utilizados para cálculos adicionales. Posteriormente, se llevó a cabo un análisis de la distribución de las variables con el fin de tener un análisis inicial de las columnas. Dentro de los resultados más llamativos de este análisis, se encuentra el comportamiento de la distribución de la edad, en este histograma se observa que en la cola izquierda se encuentran individuos con edades de -1000 o 0 años y la cola derecha personas con 115 años.

Gráfico 1. Edades de la población



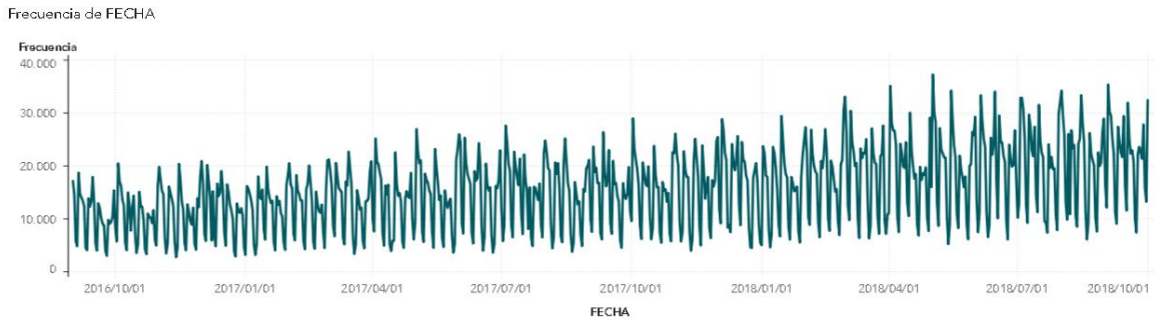
Fuente: Datos Datatón, Cálculos propios.

1.2. Análisis gráfico

Para el análisis gráfico se utilizó SAS Viya en donde se realizaron cruces de variables con el fin de encontrar patrones de comportamiento entre los gastos, días de la semana, días del mes y entre las demás variables suministradas. A continuación, se revisarán algunos casos de interés hallados en los datos.

Se inicia con la revisión de los datos mostrando la frecuencia de los gastos en todo el periodo de análisis propuesto. Se resalta el ciclo que tienen los datos relacionados con las quincenas de pagos de los usuarios, lo cual, muestra donde son los picos de gastos.

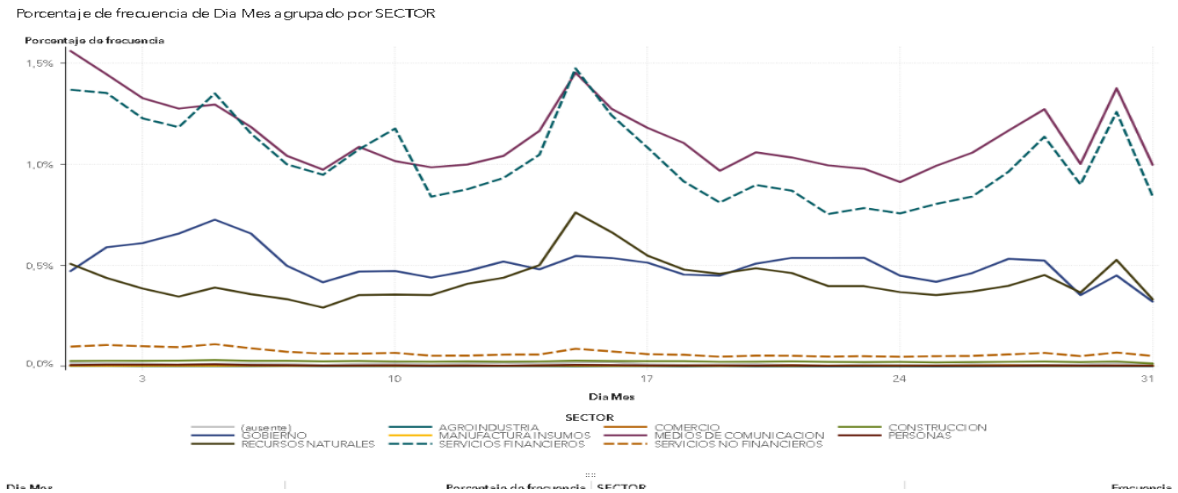
Gráfico 2. Frecuencia del gasto según la fecha



Fuente: Datos Datatón, Cálculos propios.

Primero, Se observa un claro ciclo del gasto, que presenta un pico de gasto entre los días lunes o martes (cuando lunes es festivo). También, resalta como un día importante de gasto el día de quincena o pago del salario. También, se aprecia que los sectores más importantes son Personas y servicios financieros, con claros efectos estacionales a mitad y a final del mes.

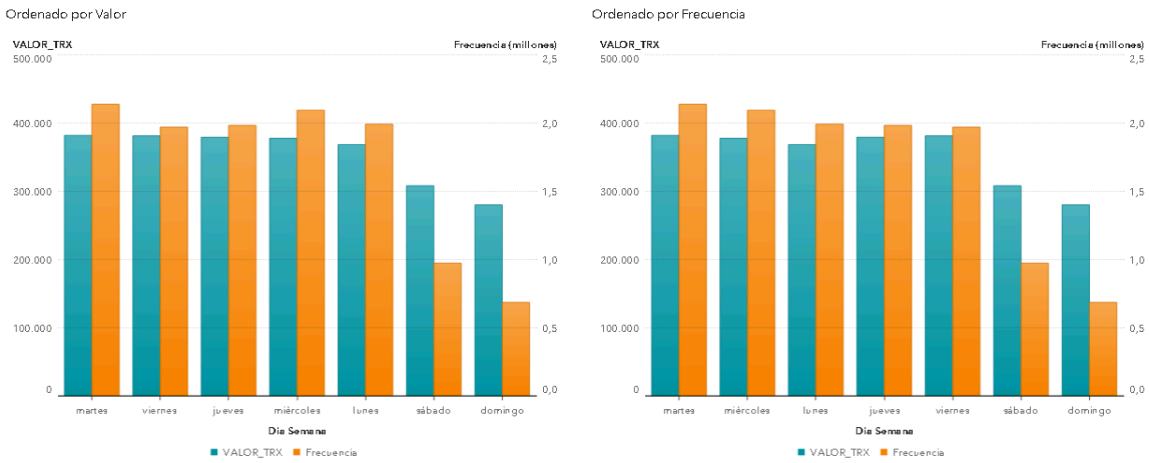
Gráfico 3. Porcentaje de frecuencia de día mes agrupado por Sector



Fuente: Datos Datatón, Cálculos propios.

En segundo lugar, se revisará el efecto estacional del consumo en los días de la semana. En el Gráfico 3, se muestran al domingo como el día de menor gasto. Los días de la semana donde mayor gasto se realiza son martes y viernes. El martes se realiza el mayor gasto monetario, seguido del viernes. Ahora, en cantidad de transacciones los días más representativos son martes, miércoles y lunes.

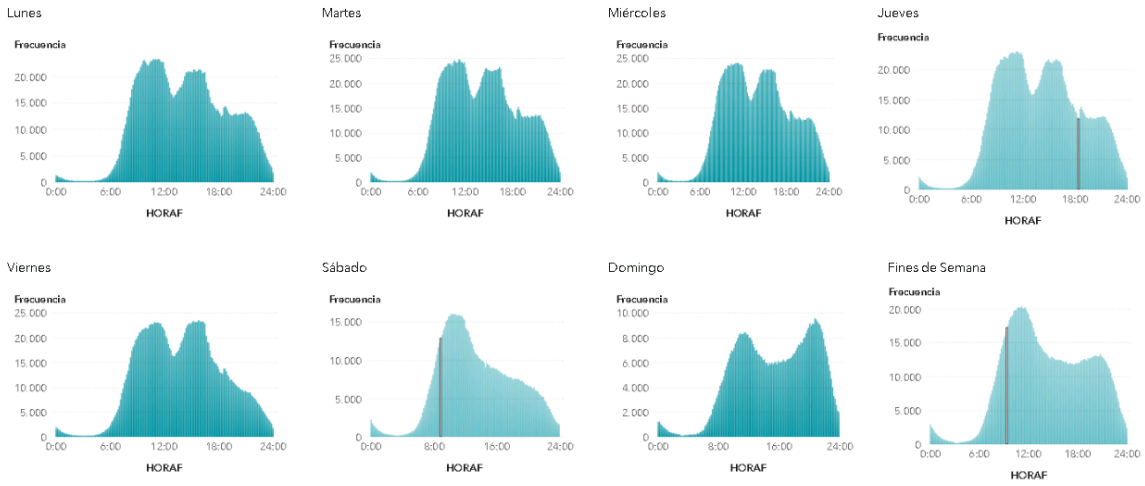
Gráfico 4. Gasto según días de la semana



Fuente: Datos Datatón, Cálculos propios.

Luego, un factor que podemos mirar asociado a los días de semana son las horas que destinan los usuarios a realizar los pagos o gastos por medio de PSE. En esta revisión, se aprecia que las horas de consumo o pagos están relacionadas con la jornada laboral, que va de las 8:30 a 12 am (mayor peso) y luego, de 2 pm a 4 pm.

Gráfico 5. Frecuencia del gasto según los días de la semana



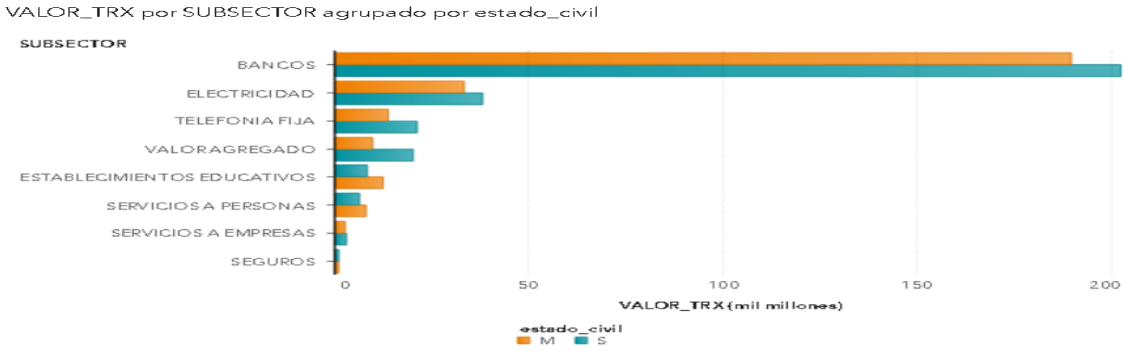
Fuente: Datos Datatón, Cálculos propios.

Las personas de mayor ingreso son las que usan los medios virtuales para realizar sus compras y pagos. Sin embargo, sigue siendo las compras por internet un ítem de segundo o tercer renglón, debido a que al parecer persiste una predominancia de los gastos en efectivo (baja penetración bancaria)

Patrones de horas: no se modifican al incluir las otras variables de los datos, debido a esto se explora si hay recurrencia en los datos dado el monto de la transacción. No existen transacciones iguales (a excepción de aquellas que son de control del “sistema”). Debido a esto se explora si pueden estar contenidas en un rango. Los resultados muestran que, al clasificar el tipo de gasto, utilizando la columna de texto ref1, no se logra generar una clasificación ya que los diferentes niveles se superponen al analizarlos frente a la variable mono de transacción.

Otro caso, muestra la relación entre los principales gastos filtrados por estado civil, según lo encontrado en los datos se observa que en las categorías establecimientos educativos y servicios de personas, como era de esperar los individuos casados (M) son los de mayores gastos. En las restantes categorías los solteros (S) son los individuos con mayores gastos, se podría suponer que los individuos con menores obligaciones tienden a gastar mucho más o a no tener control sobre los gastos.

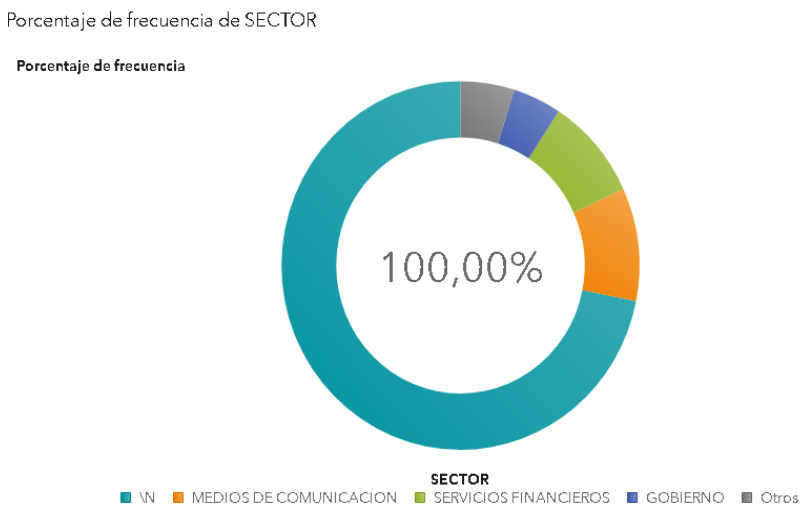
Gráfico 6. Principales gastos por estado civil



Fuente: Datos Datatón, Cálculos propios

En este mismo sentido, se revisa la frecuencia del sector Gráfico 7, esto con el fin de confirmar cuales son los más destacados. Ahora, en esta categoría destacan los valores sin ninguna clasificación o N.A que representa cerca del 71,9% de los datos, seguido de los sectores medios de comunicación y servicios financieros. A continuación, se revisan los subsectores más destacados.

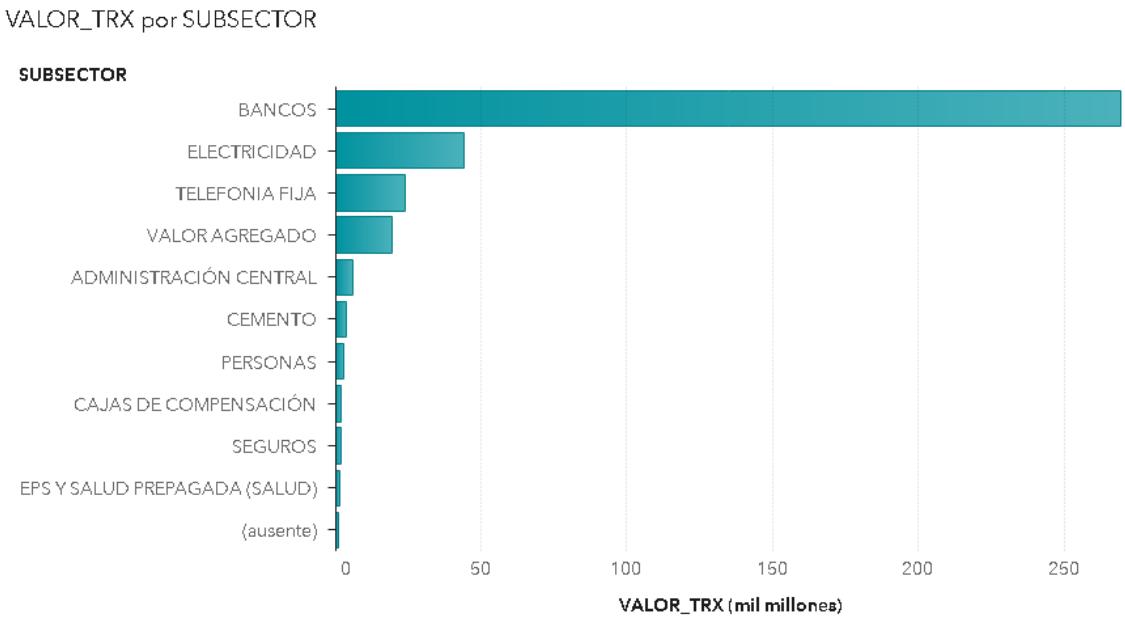
Gráfico 7. Frecuencia por sector



Fuente: Datos Datatón, Cálculos propios

En materia de los subsectores más destacados resaltan categorías como Bancos, Electricidad, Telefonía fija, Valor Agregado, Administración Central.

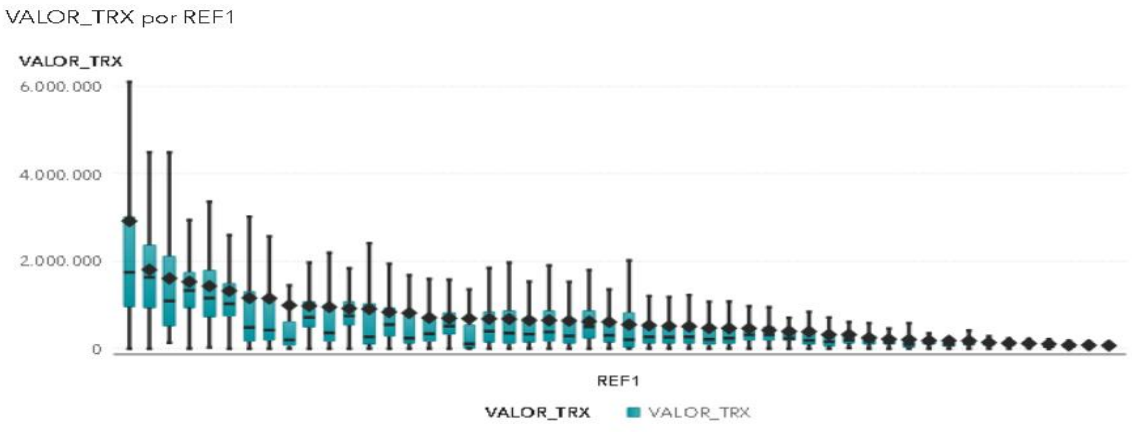
Gráfico 8. Frecuencia por subsector



Fuente: Datos Datatón, Cálculos propios

En el Gráfico 9, se pretende verificar, si las referencias tienen un rango único que permita clasificarlas en alguna categoría. En otras palabras, referencias con valores de transacción dentro del mismo rango, pertenecerían a la misma categoría. Sin embargo, como se observa en el gráfico, gran porcentaje de las 487639 referencias distintas pertenecen al mismo rango. Esto causa, que no se pueda realizar una clasificación con esta variable.

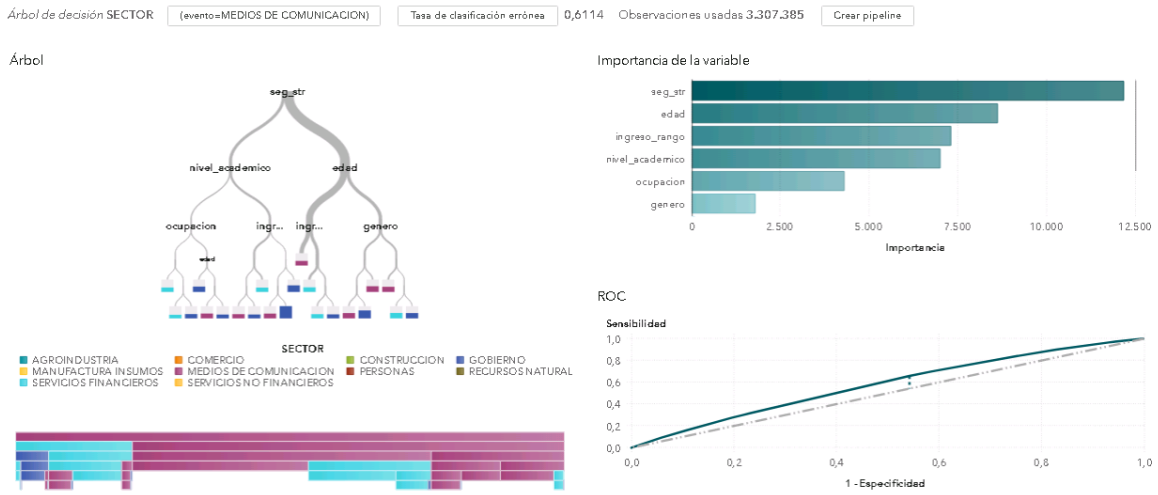
Gráfico 9. Rangos de valor transacción por referencia (ref1)



Fuente: Datos Datatón, Cálculos propios

Tomando en consideración los resultados anteriores, se procedió a tratar de clasificar los datos por medio de la variable sector. Se tomó el 28,8% de los datos, pues, solo estos cuenta con observaciones no nulas del sector. Esto con el fin de entrenar un modelo para clasificar el restante 71,9% de los datos. En este sentido, se utilizó, un modelo de machine learning, específicamente un modelo de árboles de decisiones. Entonces, los resultados de este modelo no fueron estadísticamente significativos, tal como se puede apreciar en la Gráfico 10. Por lo tanto, se optó por realizar una clasificación de la columna ref1, que es de texto libre, por lo que se necesita recurrir a la analítica de texto.

Gráfico 10. Árbol de decisión por sector.



Fuente: Datos Datatón, Cálculos propios.

1.3. Minería de texto

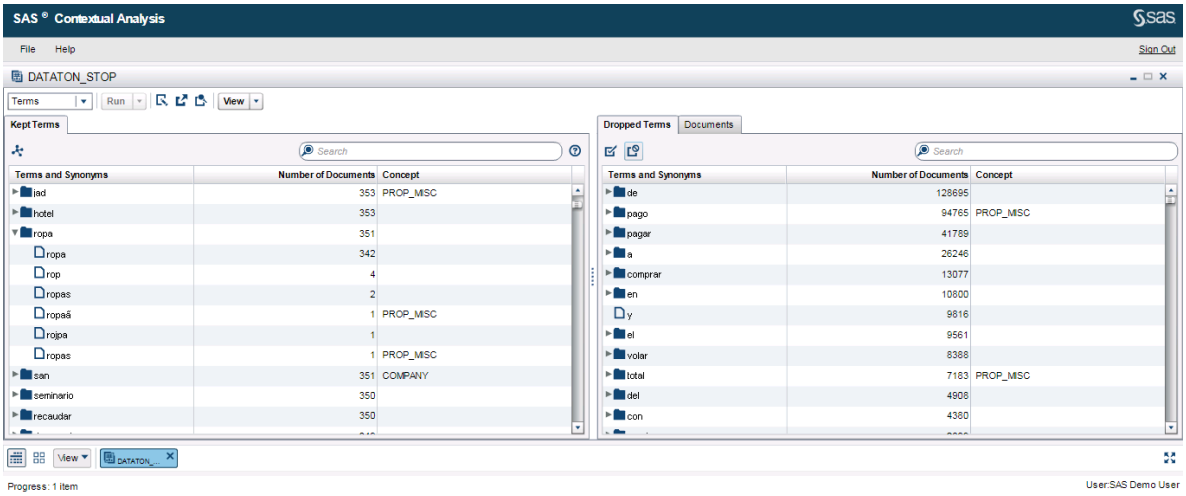
Con el fin de llevar a cabo la clasificación de las transacciones, se realizó un análisis contextual de la columna ref1 que contiene detalles de la transacción realizada por el usuario. El análisis contextual de texto se ejecutó utilizando la herramienta de SAS Contextual Analysis. Esta herramienta es usada para clasificar una gran cantidad de colecciones de documentos, visualizar, descubrir patrones en la data de texto y generar automáticamente reglas lingüísticas mejoradas con inputs generadas por la máquina o por el usuario. En este sentido, SAS Contextual Analysis permite una aproximación híbrida entre técnicas de machine learning y refinamiento realizado por el usuario.

La metodología del análisis contextual de texto toma en primer lugar, los datos de texto no estructurado, en este caso, el texto que contienen las filas de la variable ref1, y lleva a cabo un procesamiento de lenguaje natural que consiste en un análisis no supervisado con técnicas de machine learning y una imputación por el usuario.

El procesamiento de lenguaje natural brinda como resultados tópicos, análisis con valor agregado, relaciones entre palabras, taxonomías y clasificación de documentos (o filas de texto), que pueden ser usados para análisis futuros. En este caso, los resultados del análisis contextual serán utilizados para la clasificación de los gastos de los usuarios, así como para analizar las oportunidades de negocio que tiene el banco.

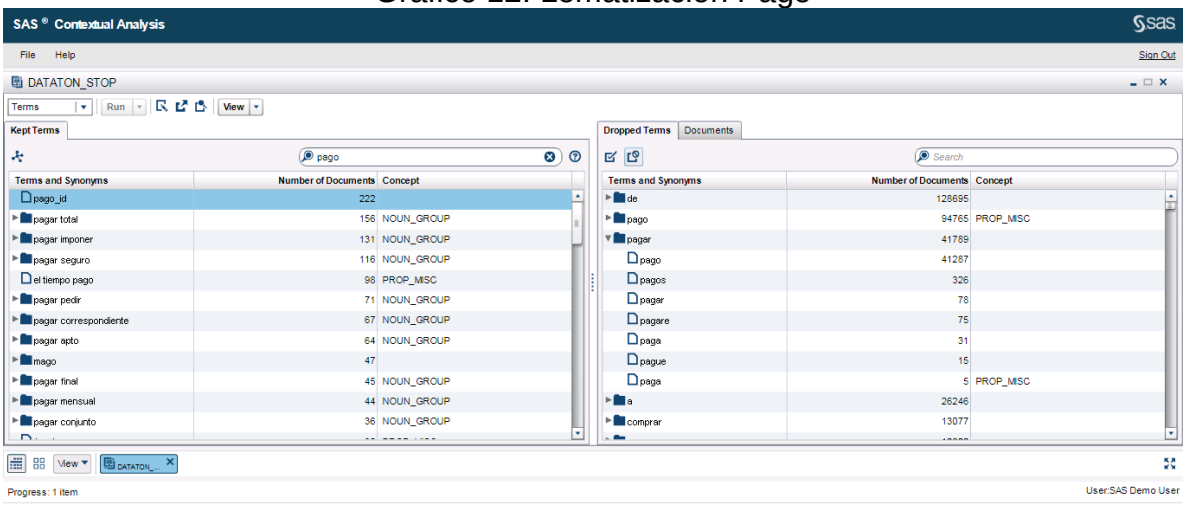
El procesamiento de lenguaje natural se divide en: análisis no supervisado con técnicas de machine learning e imputación del usuario. La primera, ayuda en el descubrimiento de los datos de texto creando agrupaciones de palabras. En este sentido, el análisis no supervisado en primer lugar lematiza o “tokeniza” de texto con el fin de agrupar sinónimos y palabras que comparten la misma estructura gramatical encontrando 40.350 términos distintos. Un ejemplo de este proceso se muestra en el Gráfico 11 y 12 en donde los términos “ropa”, “ropas”, “rojpa”,”rop” y “ropa” se lematizan en un solo término: “ropa”. Por otro lado, en la gráfica 12 se lematizan las palabras “pago”, “pagos”, “pagar”, “pagaré”, “pague”, “paga” en un solo término llamado “pagar”

Gráfico 11. Lematización Ropa



Fuente: Datos Datatón, Cálculos propios.

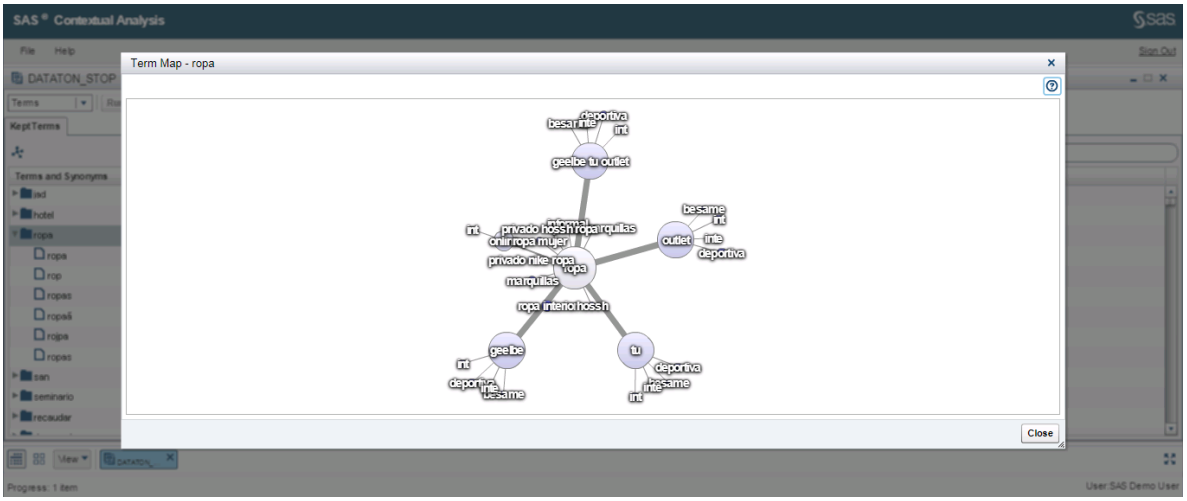
Gráfico 12. Lematización Pago



Fuente: Datos Datatón, Cálculos propios.

En este punto, es posible observar las relaciones de las palabras seleccionadas con el fin de hacer posteriormente una refinación de los términos. Un ejemplo de esto se muestra en el Gráfico 13, en donde se encuentran las palabras que se relacionan con el término “ropa” y cómo es esta relación.

Gráfico 13. Mapa de Términos Ropa



Fuente: Datos Datatón, Cálculos propios.

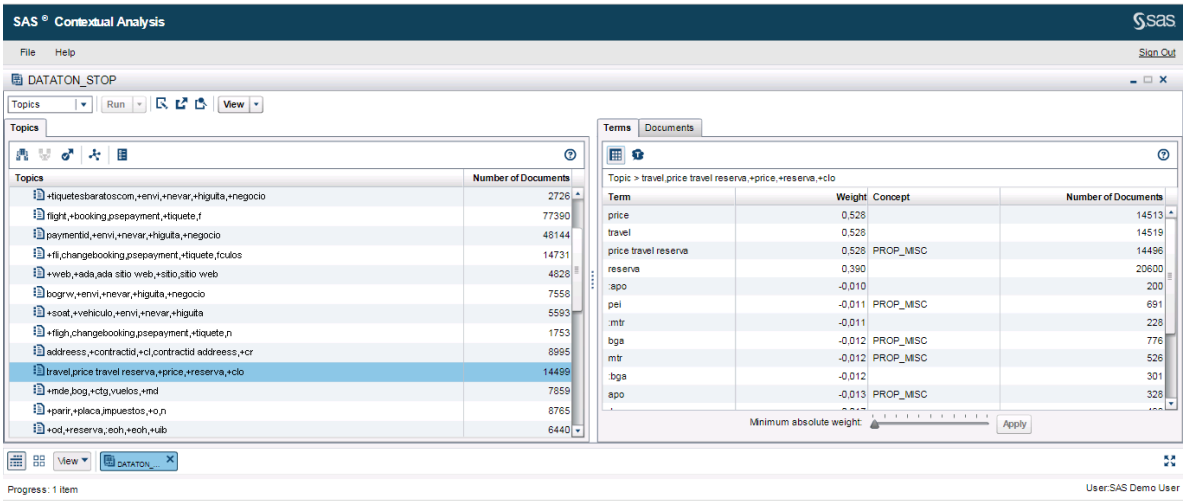
En este sentido, el término ropa se relaciona directamente con los términos “geelbe” (un outlet en línea de ropa y accesorios[1]), “outlet” y “geelbe tu outlet”, e

[1] https://www.geelbe.com/?gclid=EAIaIQobChMIqdCCndKq3gIVEF8NCh3KVwKsEAAYASAAEgJT_PD_BwE

indirectamente con palabras como “mujer”, “int”, “deportiva” y “nike” por poner algunos ejemplos.

Una vez se ha lematizado el texto, el análisis no supervisado asigna a cada uno de los términos (palabras o términos compuestos) que compone el texto, un peso tal que permite la posterior agrupación de estos términos en un clúster, que al ser unidos, generan temas tal y como se puede observar en la Gráfica 14, que muestra un caso particular con los términos relacionados con viajes.

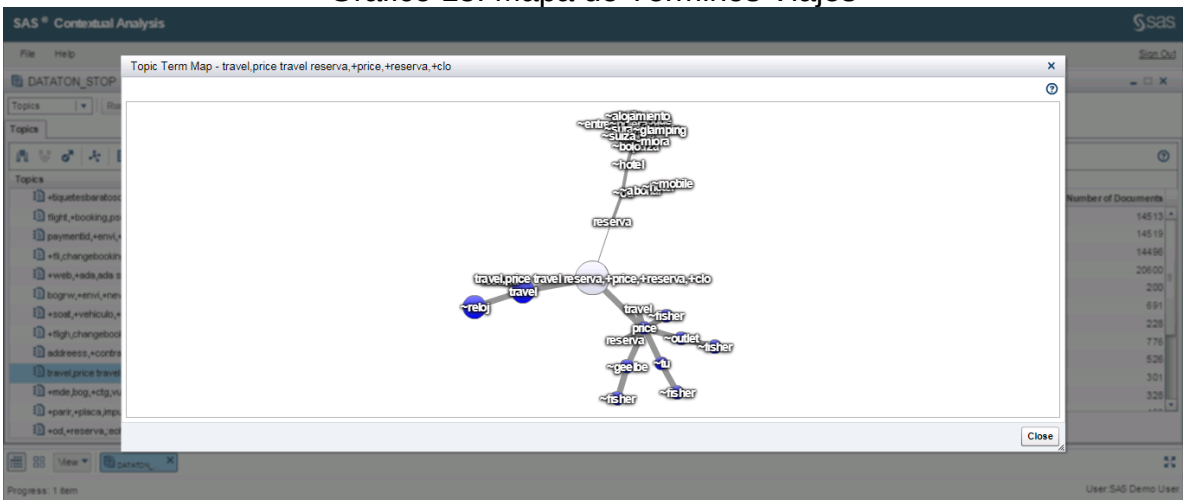
Gráfico 14. Tópico Viaje



Fuente: Datos Datatón, Cálculos propios.

Adicionalmente, en este punto se puede volver a explorar con qué términos y de qué manera estos están relacionados con el tópico tal y como lo muestra el Gráfico 15.

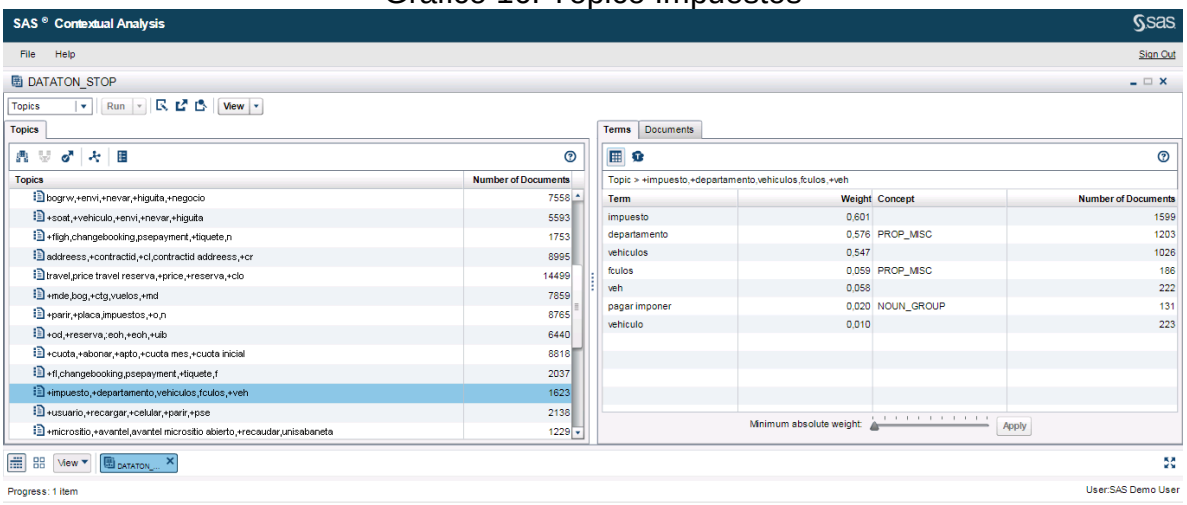
Gráfico 15. Mapa de Términos Viajes



Fuente: Datos Datatón, Cálculos propios.

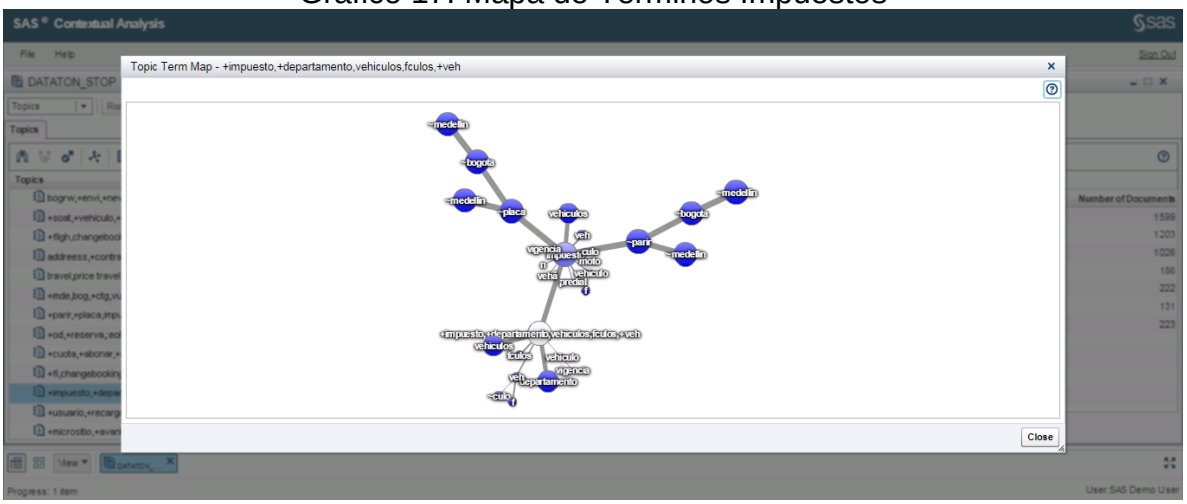
En el gráfico anterior, se observa el tópico que hace referencia a tiquetes de avión se relaciona directamente con términos como “travel”, “reserva” y “Price” e indirectamente, por ejemplificar, con términos como “hotel” y “alojamiento”. Otro ejemplo de los análisis que se pueden desarrollar con esta herramienta que permitirán la refinación de estos tópicos para la construcción de conceptos, es el que se realizó con el tópico que hace referencia al pago de impuestos y que se muestra en el gráfico 16 y 17.

Gráfico 16. Tópico Impuestos



Fuente: Datos Datatón, Cálculos propios.

Gráfico 17. Mapa de Términos Impuestos

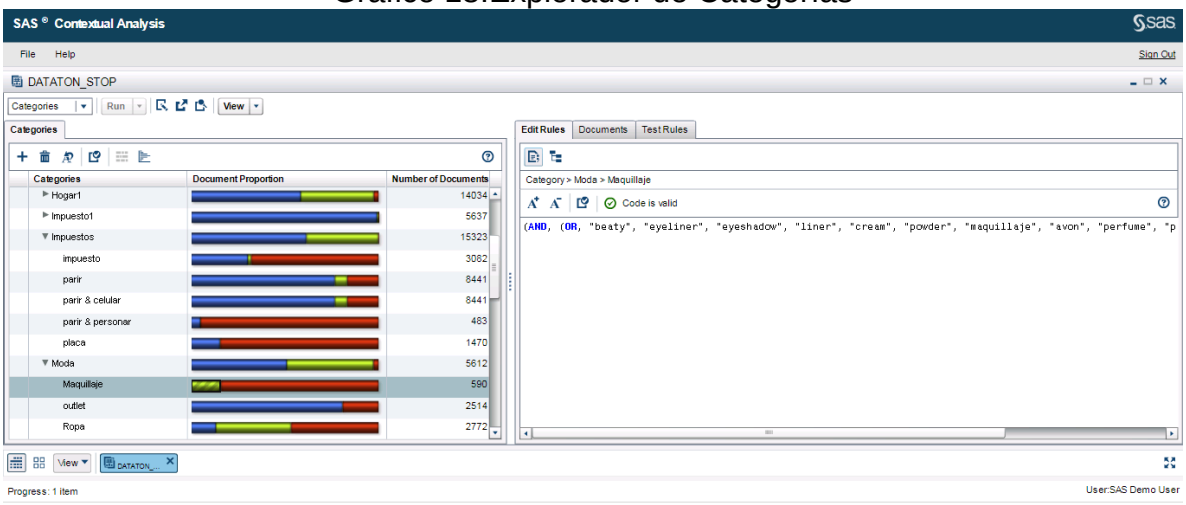


Fuente: Datos Datatón, Cálculos propios.

En estos gráficos podemos observar como el tópico lo componen términos como “impuesto”, “departamental”, “vehículo” y “pagar imponer” y se relaciona con términos como “vehículo”, “moto” y “departamento”.

Una vez este proceso termina, se hace el refinamiento manual de los temas uniendo y separando clústers a partir de los insights obtenidos con los análisis realizados en la etapa de tópicos. Adicionalmente, SAS Contextual Analisys permite incorporar el conocimiento de expertos a través de la creación manual de tópicos, que se componen de términos seleccionados por este experto, y el renombramiento de los que fueron creados automáticamente. Un ejemplo de esto lo podemos observar en el Gráfico 18 y en el Gráfico 19.

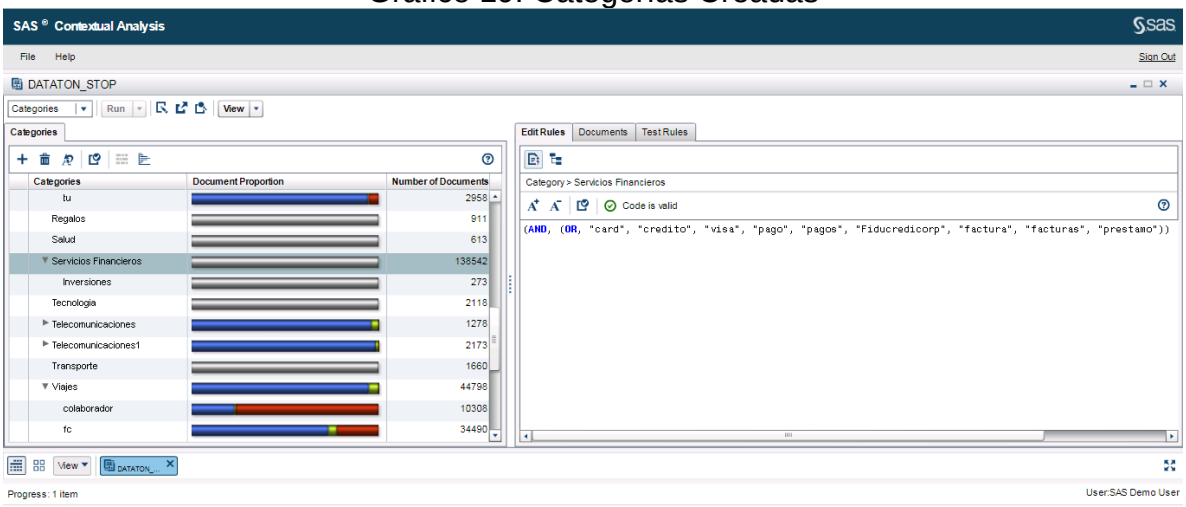
Gráfico 18.Explorador de Categorías



Fuente: Datos Datatón, Cálculos propios.

En el Gráfico 19, podemos observar cómo el tópico de Moda, puede ser segmentado en varios sub-tópicos tales como “Maquillaje”, “outlet” y “ropa”. Además, en la parte derecha de la pantalla, los términos de los que se compone este tópico. Finalmente, la herramienta clasifica el texto y muestra el desempeño de esta clasificación por tópico. En este sentido, la barra azul muestra el porcentaje de verdaderos/positivos, el rojo el porcentaje de falsos/negativos y el verde los falsos positivos lo que permite un proceso de afinación mayor de la categoría.

Gráfico 19. Categorías Creadas



Fuente: Datos Datatón, Cálculos propios.

Por otra parte, el Gráfico 19, muestra un ejemplo de las categorías creadas manualmente tal como la categoría “Servicios Financieros” que se compone de los términos “card”, “credito”, “pago”, “factura” entre otros y que tiene una subcategoría denominada inversiones que tiene palabras como “btc” y “bitcoin”. Cabe resaltar que las barras de las categorías creadas manualmente son grises, ya que no poseen las tasas de verdadero/positivo, falso/negativo y falso/positivo.

Con estas categorías ya creadas, se procedió a hacer un scoring y clasificar cada una de las transacciones de los datos en las categorías encontradas, en este caso fueron 12 categorías con subcategorías en donde 6 categorías fueron creadas manualmente.

2. Resultados de aplicar Analítica de Texto

Una vez finalizada la analítica de texto, se logra llegar a la siguiente tabla en donde se muestra la frecuencia de cada una de las categorías y subcategorías encontradas, así como el monto total por cada una de ellas.

Tabla 1. Resultados de la analítica de texto

	N	VALOR_TRX
		Sum
name		
Top/Deporte	5,181.00	2,858,910,432.10
Top/Educacion	62,896.00	59,636,721,483.00
Top/Electrodomesticos	127,134.00	39,436,479,779.00

Top/Electrodomesticos/alkosto	33,611.00	13,086,467,134.00
Top/Electrodomesticos/linear	90,708.00	24,982,586,758.00
Top/Electrodomesticos/n & x & cm	3.00	98,009.28
Top/Electrodomesticos/tv & cm	936.00	1,311,008,135.80
Top/Electrodomesticos/x & cm	1,876.00	56,319,741.64
Top/Entretenimiento	19,250.00	12,242,295,972.00
Top/Hogar	451,419.00	95,951,738,105.00
Top/Hogar1	699,287.00	217,254,602,624.00
Top/Hogar1/Vivienda	330,591.00	99,238,354,843.00
Top/Hogar1/abonar & agosto	133.00	137,138,066.97
Top/Hogar1/abonar & apartamento	234.00	624,176,114.96
Top/Hogar1/abonar & apto	505.00	1,569,957,458.60
Top/Hogar1/abonar & credito	1,292.00	865,191,181.94
Top/Hogar1/abonar & enero	135.00	207,892,283.92
Top/Hogar1/abonar & julio	165.00	205,697,254.05
Top/Hogar1/abonar & junio	177.00	356,072,703.00
Top/Hogar1/abonar & septiembre	137.00	168,238,957.28
Top/Hogar1/apto & administraci	248.00	50,506,166.70
Top/Hogar1/contrato	304,882.00	70,963,274,140.00
Top/Hogar1/cuota	60,788.00	42,868,103,453.00
Top/Impuesto1	33,929.00	30,197,406,158.00
Top/Impuesto1/soat	33,929.00	30,197,406,158.00
Top/Impuestos	236,154.00	113,366,744,333.00
Top/Impuestos/impuesto	85,134.00	26,780,902,424.00
Top/Impuestos/parir	36,371.00	8,751,811,467.20

Top/Impuestos/parir & celular	36,371.00	8,751,811,467.20
Top/Impuestos/parir & personar	3,538.00	272,976,281.66
Top/Impuestos/placa	5,645.00	725,163,960.00
Top/Moda	24,194.00	2,605,301,307.70
Top/Moda/Maquillaje	2,158.00	288,342,532.18
Top/Moda/Ropa	5,362.00	649,021,293.81
Top/Moda/outlet	9,551.00	939,414,027.15
Top/Moda/tu	7,123.00	728,523,454.53
Top/Regalos	1,895.00	250,237,802.35
Top/Salud	48,144.00	24,300,910,475.00
Top/Servicios Financieros	6,105,011.00	1,947,899,700,000.00
Top/Servicios Financieros/Inversiones	7,591.00	2,841,868,508.10
Top/Tecnologia	13,070.00	1,500,430,643.20
Top/Telecomunicaciones	5,406.00	2,015,115,710.50
Top/Telecomunicaciones/avantel	1,577.00	97,639,348.25
Top/Telecomunicaciones/micrositio	3,794.00	1,855,320,617.90
Top/Telecomunicaciones/unisabaneta	35.00	62,155,744.31
Top/Telecomunicaciones1	630,436.00	81,266,315,998.00
Top/Telecomunicaciones1/celular	11,594.00	855,070,327.61
Top/Telecomunicaciones1/recargar	599,549.00	61,910,304,820.00
Top/Telecomunicaciones1/usuario	19,293.00	18,500,940,851.00
Top/Transporte	346,084.00	12,778,964,554.00
Top/Viajes	53,521.00	13,368,001,205.00
Top/Viajes/colaborador	10,520.00	1,342,057,895.20
Top/Viajes/fc	43,001.00	12,025,943,310.00

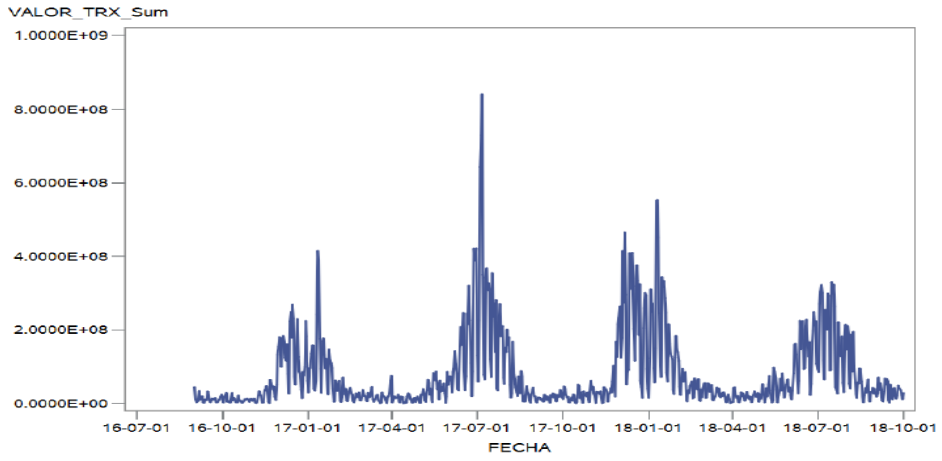
Top/Viajes1	2,110,348.00	495,269,019,576.00
Top/Viajes1/Payment	1,947,156.00	450,751,221,529.00
Top/Viajes1/booking	88,218.00	24,316,796,989.00
Top/Viajes1/fligh	1,896.00	344,740,142.48
Top/Viajes1/flight & f	68,934.00	19,097,696,871.00
Top/Viajes1/tiquete & fl	4,144.00	758,564,045.24
Top/Viajes3	58,263.00	18,585,600,744.00
Top/Viajes3/ada	4,406.00	1,286,473,613.30
Top/Viajes3/sitio	53,857.00	17,299,127,131.00
Top/Viajes4	60,044.00	28,676,951,052.00
Top/Viajes4/bog	20,264.00	10,784,411,295.00
Top/Viajes4/bog & baq	1,556.00	656,073,399.39
Top/Viajes4/bog & bga	1,085.00	367,100,911.27
Top/Viajes4/bog & clo	2,226.00	1,134,516,026.50
Top/Viajes4/bog & ctg	2,009.00	946,235,392.06
Top/Viajes4/bog & cuc	877.00	399,885,465.11
Top/Viajes4/bog & mde	4,317.00	2,121,255,534.00
Top/Viajes4/bog & mtr	611.00	274,082,797.92
Top/Viajes4/bog & od	882.00	204,898,040.98
Top/Viajes4/bog & pei	1,098.00	421,179,329.75
Top/Viajes4/bog & smr	1,197.00	624,991,075.11
Top/Viajes4/fli	6,432.00	1,212,676,964.90
Top/Viajes4/mde & clo	534.00	320,672,232.15
Top/Viajes4/mde & ctg	791.00	334,308,209.43
Top/Viajes4/price	16,165.00	8,874,664,378.60

Top/Viajes5	15,966.00	4,830,821,695.80
Top/Viajes5/bogrw	7,559.00	2,410,854,231.20
Top/Viajes5/flig	5,627.00	1,011,761,563.60
Top/Viajes5/tiquetesbaratoscom	2,780.00	1,408,205,901.10
Total (ALL)	15,100,810.00	4,185,922,400,000.00

En la tabla anterior, se logra observar como la categoría con mayor cantidad de transacciones de servicios financieros 40%, seguido de la categoría de viajes con un peso del 35% y, en tercer lugar, con el 10% del total de transacciones se encuentra la categoría de hogar. Cabe resaltar que la incongruencia entre los 12 millones de transacciones originales y los 15 millones de transacciones que se muestran en la tabla anterior es generada debido a que varias transacciones pueden caer dentro de varias categorías lo que impide que se hagan también análisis sobre los montos.

Siguiendo, con el análisis de los resultados, a partir de la clasificación de las transacciones se logran crear los gráficos del valor de la transacción total por fecha, categoría y sub-categoría que pueden ser encontrados en el Anexo 1. A continuación en el Gráfico 20 y 21, se muestra el análisis gráfico del monto de transacciones para el sector de educación y de impuestos de vehículos respectivamente.

Gráfico 21. Valor de transacción por fecha sector educación

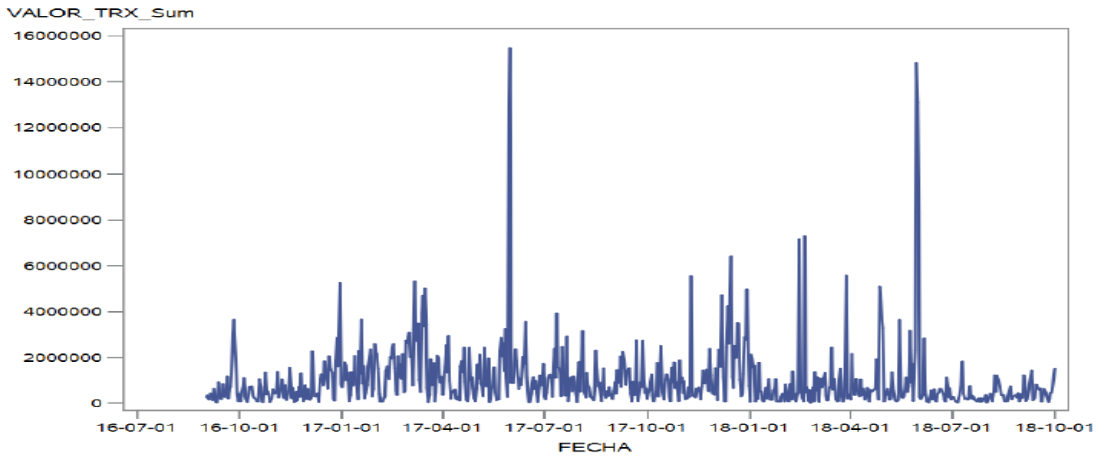


Fuente: Datos Datatón, Cálculos propios.

EL gráfico 21, muestra una periodicidad en las transacciones relacionadas con educación, específicamente se puede observar como entre los meses de Noviembre-Diciembre y Junio-Julio existe un incremento de las transacciones.

Mientras, que entre Enero-Febrero y Julio-Agosto, se observan disminuciones en los pagos de este rubro alcanzando los picos en Enero y Julio, lo que es consistente con los periodos de cambio de semestres de las universidades tanto a nivel de pregrado y de posgrado. Este resultado, valida la eficiencia del proceso de clasificación implementado.

Gráfico 22. Valor de transacción por fecha sector impuestos vehiculares



Fuente: Datos Datatón, Cálculos propios.

Por otra parte, el gráfico 22, muestra también una periodicidad con las transacciones de impuestos vehiculares. En primer lugar, se nota como estas transacciones tienen picos entre abril y Julio, que coinciden con las fechas límite para el pago de este impuesto sin multa. Adicionalmente, se observa como antes de Abril existen picos pero menores a los mencionados anteriormente que coinciden también con las fechas límite para pagar este rubro con descuento.

3. Conclusiones

El proceso de análisis de los datos nos permitió identificar que los intervalos de gasto por el campo ref1 no tiene una clasificación excluyente, lo que no permite identificar un patrón haciendo uso de las variables socioeconómicas. Sumado a esto, se procedió a tratar de clasificar los datos por medio de la variable Sector, a través de un modelo de árboles de decisión, sin embargo, este modelo no es estadísticamente significativo. Lo que rebela que no existen patrones o relaciones identificables entre las variables socioeconómicas, como para utilizarlas para realizar una clasificación de los datos. Esto es debió a que los usuarios realizan sus pagos por diferentes canales, de tal manera, que si quisiéramos encontrar recurrencia en los pagos se necesita información del POS, cajeros automáticos,

PSE, etc. Otro punto a destacar para el proceso de educación financiera es poder contar con una mayor precisión en el rango de ingresos, para saber si el cliente esta gastando más de su presupuesto, es decir, se esta endeudando y poder brindarle apoyo con educación financiera y oferta de productos. En síntesis, se tomó la decisión de usar analítica de texto en el campo ref1 para realizar el proceso de clasificación.

La utilización de los modelos de minería de texto implica dedicar una gran cantidad de tiempo a la creación de diccionarios y afinar más las categorías para tener mayor exactitud en la clasificación. La propuesta metodológica permite realizar una categorización para realizar el push para el envío de mensajes que apoyen las estrategias de finanzas personales.

La propuesta de valor es la manera en la que estamos identificando y realizando las diferentes categorías con una metodología de machine learning para analítica de texto, lo que permite clasificar toda la base de datos y adicional todas las transacciones que vayan llegando. Este proceso, se puede mejorar por medio de una afinación de los términos más significativos