EQUIPO NYIA

Clasificación de transacciones y segmentación de clientes a partir de hábitos de pagos vía PSE

Olga Escobar
Laura Caro
Alexis CArrillo

Objetivos:

En el desarrollo de este reto, como equipo nos planteamos do objetivos:

- Realizar una clusterización de clientes, con de conocer el comportamiento de estos a la hora de realizar transacciones utilizando el botón de pago PSE.
- Utilizar un método de procesamiento de lenguaje natural que permita predecir el sector a partir de la ref1

Metodología

Durante el desarrollo se llevaron a cabo los siguientes pasos

1. Exploración y limpieza de los datos
2. Elección de los modelos a implementar, en este caso fueron dos:
    - Para una muestra del 1.5% se realizó Dendrograma de Ward y el resto de los datos se asignaron a él clúster con clasificación de KNN
    - Se va a utilizar el Modelo Pipeline: Spark Machine Learning Peipelines Api, el cual es similar a Scikit _ Learn, para desarrollar el modelo de predicción

Resultados

Los resultados obtenidos muestran en los datos ***dt_info_pagadores_muestra.csv (****lo llame* ***pagadores)*** las siguientes particularidades para estas variables:

- tipo_vivienda tiene el 50.8 % de datos nulos
- ocupación 2.01% de datos nulos
- nivel_academico 13.08%
- estado_civil 1.94%
- genero 1.68%

Considero quitar la variable tipo_vivienda y quitar todos los NA

De los datos de **dt_trxpse_personas_2016_2018_muestra_adjt.csv (**lo llamé **personas)** se tiene:

Variable	% Datos Nulos
id_trn_ach	0.00037
id_cliente	0.0053
fecha	0.107
hora	0.00503
valor_trx	0.0053
ref1	3.0524
ref2	42.79
ref3	100
sector	0.209
subsector	0.209
descripcion	0.209

Considero quitar las variables ref2 , ref3 y descripción, esta última utilizarla para realizar un text mining por separado, con el fin de conocer las palabras más utilizadas

Preprocesamiento

Base de Transacciones

La base de transacciones se procesó con 11 variables y 11839512 registros. Se diseñó un loop que revisaba si el registro era válido (los 12 valores) o si tenía uno de más qe era debido a una coma mal interpretada como separador.

Para las variables se hicieron las siguientes manipulaciones:
Fecha y hora se integraron para transformarlas a formate datetime.
Referencia 1, 2 y 3 se unieron en una sola referencia

El valor de la transacción se tomó como la variable principal y se exploró para identificar outliars. Se transformó el valor de la transacción en percentiles y se tomó en cuenta su orden respecto a los otros valores de transacciones. El cruce de estas variables permitió identificar un efecto techo y a la vez un cambio inusual en el comportamiento de la variable sobre los 16 millones, y teniendo en cuenta la documentación de bancolombia y la de el portal PSE, se definió este valor como el límite superior de las transacciones. Para el límite inferior se tomaron en cuenta los mismos factores y se encontró un efecto escalonado y de piso que dió lugar a la decisión de establecer lo 950 pesos como el límite inferior de las transacciones.
Se eliminaron 16927 transacciones inferiores a 950 y 5925 transacciones superiores a 16 millones, dejando en total 11816660 transacciones.

Luego del filtrado de los outliars se procedió a obtener nuevas variables para los algoritmos de machine learning
Para las transacciones se agregaron la hora del día, el día de la semana, el día del mes el mes del año y la semana del año.

Base de Personas

De la base de transacciones se resalta que la variable edad fue explorada y según la distribución se decidió eliminar las cuentas de menores de 5 años y las de mayores de 97.

Bases combinadas:

Luego de obtener los referentes temporales se procesaron para cada uno de los clientes tanto el conteo como por suma total de transacciones por día de la semana, hora del día, día del mes, y semanas del año, ampliando la base de datos demográficos a los valores propios de las transacciones.

Predicción de sector a partir de las referencias y las transacciones anotadas con el sector:

Clustering de los clientes según variables demográficas y transacciones de PSE:

Se tomaron las variables integradas que dan información acerca de la frecuencia y la cantidad de dinero por marcas temporales para cada uno de los clientes. Estos valores numéricos se normalizaron con percentil, ya que hace balanceada la distribución de la variable y permite calcular distancias de manera más homogénea, sin ser afectado por el sesgo de la distribución de la variable..

PSE Pagos Seguros en Línea, tiene muchas ventajas y beneficios paras empresas financieras como el banco Bancolombia, dado que las transacciones se realizan en tiempo real, consolida automáticamente la información en línea, ahorra gastos funcionales, evita errores en pagos, acceso a un gran número de cuentas en diferentes entidades bancarias, van en aumento los niveles de recaudo y han descongestionado los centros de atención.

En las entidades bancarias se caracterizan por el nivel de incertidumbre en sus portafolios estos se ven afectados según el comportamiento del mercado y la calidad crediticia de los deudores. Este nivel disminuye a medida que se logre predecir el comportamiento de determinada variable en el mundo financiero. Para este fin se usará la clasificación Regresión Logística.

La regresión Logit se utiliza cuando queremos predecir un resultado binario, por ejemplo, quiebra vs. no quiebra y sabemos que existen varios factores que pueden incidir sobre tal resultado. Esta regresión binaria es un tipo de análisis de regresión donde la variable dependiente es una variable dummy: código 0 (Buen Cliente) o 1 (Mal Cliente).
En general, se conoce el par (x,y), para predecir z.
Se asigna y=1 o 0.
Feacture vector x=(1,count"medios de comunicación",count"servicios financieros", count"Gobierno")
βes el conjunto de los pesos de cada parámetro.
El modelo de regresión logística es:

$$P(y = 1/x, \beta) = g(\beta^t x)$$

$$g(z) = \frac{exp(z)}{1 + exp(z)}$$

Se busca predecir a qué sector de la economía determinado cliente realizará la próxima transacción. Por medio de esta información se puede detectar cual es el sector con mayor demanda y de esta forma crear o producir un producto de alto consumo.