Documentación "Clasifica tu PyME"

Equipo: Especialistas en TI

Octubre 2020

1. Introducción

Este entregable corresponde a la Hackathon BBVA 2020, para el reto de Crédito PyME. Nuestro entregabe esta constituido por una platafroma web interactiva en la que un usuario puede consultar información relevante sobre PyME's actuales registradas, asi como la toma de decision de si tal PyME es apta o no de recibir un credito.

Los modelos que proponemos para conllevar la toma de decisiones es un modelo logistico y un modelo con grafos.

A continuacion se explican los modelos y la metodologia del procesamiento de datos.

2. Diagrama del proyecto

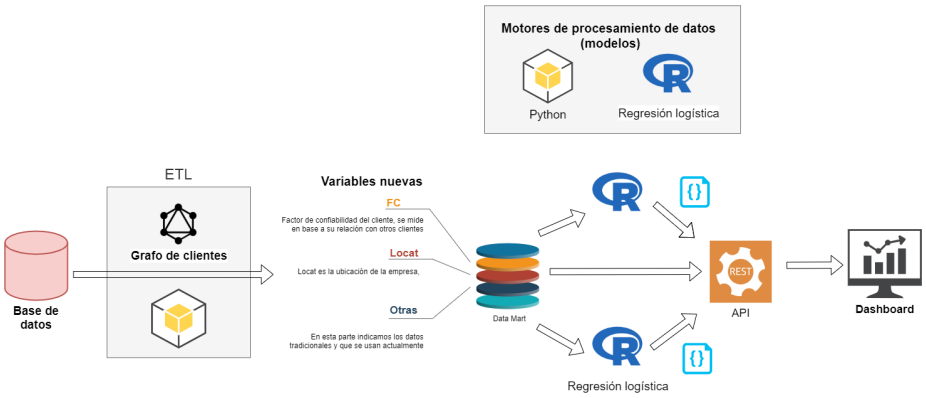

Figura 1: **Esquema general del proyecto**

3. ETL

En esta sección se realiza la extración, transformación y carga (ETL) tradicional, adicionalmente añadimos algunas variables que crearemos en base al conocimiento del negocio y factores externos que eventualmente seran usados por los motores de procesamiento de datos.

3.1. Factor de confiabilidad del cliente (FC)

La variables FC busca cuantificar cuan confiable es un cliente tomando en cuenta su circulo de contactos en primer grado.
Para ellos los datos seran representado en un grafo, donde los **nodos** seran los *clientes* y las **aristas** representaran la **relación de confianza** con los clientes.
Sea un cliente C resultado de una función de composición de las trasacciones

Algorithm 1: Fc del cliente C

Result: FC
Contactos del Cliente C;
total = [];
for *cada c* ***in*** *C* **do**
    cmp = Calcular relación de C con c;
    añadir(total, cmp)
end
FC = composición(total)

Por ejemplo, para el cliente 01238624 temenos el siguiente subgrafo de contactos.

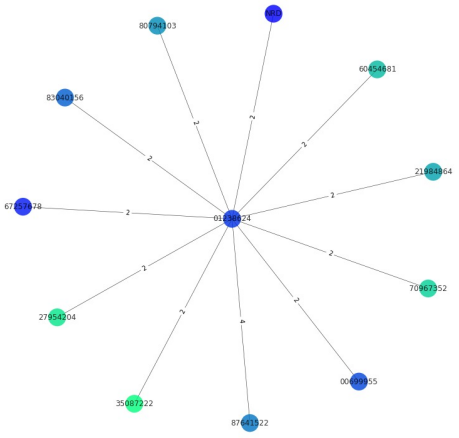

Figura 2: **Contactos del cliente 01238624**

4. Tratamiento de datos

4.1. Modelo logístico

Usando la base de datos dbRcc, hacemos uso de las siguientes variales: $sbs\_customer\_id, producto, balance\_amount, situacion\_credito$.

1. *sbs_customer_id*: Identifica a cada PyME
2. *producto*: Tipos de productos que tiene la PyME
3. *balance_amount*: Saldos que tiene la PyME en cada uno de sus productos
4. *situacion_credito*: Situación de la PyME en que se encuentra su producto, ésta puede ser VENCIDO, REESTRUCTURADO, REFINANCIADO y JUDICIAL.
5. *num_days_default_payment_number*: Numero de días en los que la PyME se ha retrasado en pagar en alguno de sus porductos en cieto sistema bancario.

Ahora bien, cada renglon de la base identifica a la PyME en cierta fecha e identidad financiera por lo que una PyME puede aparecer más de una vez. Dada las $7,962,068$ y la repetición de las PyME, reducimos el numero de renglones A 435,222 haciendo una agrupacion por id de cada Pyme, por lo que será necesario considerar nuevas variables.

Creamos nuevas variables a partir de las anteriores establecidas:

1. Con la variables *producto* creamos *count_producto*: mide el numero de prodcutos distintos que tiene una PyME.
2. Con la variable $num_days_default_payment_number$ creamos $mean_days_default_payment$: da el promedio de días en que la PyME se ha retrasado en pagar por todos sus productos.
3. Con la variable *situacion_credito* creamos *status*: Indica con un 1 si la PyME ha estado en algun momento en sus productos con estado: VENCIDO, REESTRUCTURADO, JUDICIAL O REFINANCIADO y es 0 de otro modo. En otras palabras, cuando vale 1 es indicio de que la PyME ha dejado de pagar en algun momento.
4. Con la variable $balance_amount$ creamos otra con el mismo nombre pero esta representa la media de saldos que tiene la PyME por todos sus productos.

El modelo es:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}}$$

Ajustando el modelo con ayuda de RStudio, ver códiGo fuente en repositorio, tenemos:

1. $\beta_0 = intercepto$ = -4.633
2. $\beta_1 = balance\_amount \approx 0$
3. $\beta_2 = mean\_days\_default\_payment$= 0.3733
4. $\beta_3 = count\_producto$ = 0.6324

Al final, este modelo sólo expresa la probabilidad en la que una PyME puede que deje de pagar en algun momento, sin embargo no nos determina si sí o no es apta para permitirsele algun crédito.

Para responder lo anterior, podriamos definir una *regla de dedo* en la que si una PyME tiene una probabilidad de 0.8 o más, de que en algun momento deje de pagar, entonces no es apta pára un nuevo crédito, de otro modo se le podria considerar.

Las probabilidades se pueden visualizar en la demo en el dashboard.

Comentarios: El modelo pudo haber sido multinomial considerando las 5 etiquetas en las que la situacion crediticia de sus productos se encuentra sin embargo por falta de tiempo no se logro ajustar ese mismo.

5. Método de los K-Nearest Neighbors

En está sección se explica y se construye el modelo de KNN. El conjunto de datos utilizados en este modelo es "dbRcc.csv". Para la construcción se utilizaron las siguientes 4 entradas.

1. La matriz que contiene los predictores asociada con los datos de entrenamiento, se utilizó la variable **DataTrain**.
2. La matriz que contiene los predictores asociada con los datos de prueba, se uso la variable **DataTest**.
3. La variable **NewD$status** contiene las etiquetas para las observaciones de entrenamiento y prueba.
4. Para el número de vecinos más cercanos, se tomó el valor de **k=3**.

La variable "NewD" se utilizó la función **cbid** para vincular las variables **status**, **mean_days_default_payment**, **balance_amount**,**count_producto**, **count_subproducto** y poder contruir las matrices **DataTest** y **DataTrain**.

Se tomaron valores binarios, podemos considerar dos clases "0" y ‘'1". Se construyó el modelo para predecir si han dejado de pagar o no el producto. En la matriz de confunsición, para este modelo tiene etiquetas en sus renglones con las clases reales, y sus columnas, con las predichas por el modelo.

Para analizar la calidad del modelo la matriz de confusión, es de la forma:

	0	1
0	107803	4604
1	1171	16989

La interprestación de la diagonal principal es: a)" 1" han dejado de pagar y es verdadero, b)"0" son los que han pagado y es verdadero. La otra diagonal refleja los errores del clasificador, es decir, los falsos positivos (dice que dejaron de pagar "1", pero en realidad no) o para falsos negativos (dicen que pagaron "0", pero en realidad no). Para evaluar el modelo se calcula el "accuracy", es:

accuracy=(Predicciones correctas)/(Número total de Predicciones) = 0.96.

Vemos que el modelo tiene una precisión alta, es decir, es un buen modelo.