Proyecto 2

Ciencia de datos y sus aplicaciones

Pablo Herrera

2022-10-05

Contexto

- En este informe se detalla la aplicación del algoritmo k means. Con este se construye un modelo de segmentación (*clustering*) de clientes bancarios.
- Se cuenta con una base de datos de más de 45 mil clientes de un banco. La base incluye información financiera y de sus interacciones con ejecutivos del banco. Sin embargo, también contiene información personal de los usuarios, como su estado civil, actividad laboral o el nivel educacional.
- Se define como objetivo de negocio para este análisis, identificar los riesgos que pueden presentar los clientes y anticiparlos para mantener una relación de largo plazo con estos.
- Luego de la obtención de resultados se entrega una propuesta de estrategia comercial para uno de los segmentos identificados.

I. Desarrollo del modelo de segmentación

I.1. Datos entregados

Originalmente la base de datos tiene 12 variables:

- **IdCliente:** identificador de cada cliente.
- **Fecha Nacimiento.**
- **Actividad Laboral** (categórica): detalla si el cliente es empresario, gerente, jubilado, obrero, técnico, trabajador dependiente o independiente.
- **Estado Civil** (categórica): detalla si el cliente es casado, divorciado o soltero.
- **Nivel Educacional** (categórica): detalla el máximo nivel educacional alcanzado por el cliente, es decir si cuenta con educación básica, media, técnico profesional o universitaria.
- **Tiene Mora** (binaria): reconoce si el cliente tiene mora con el banco.
- **Saldo Medio Anual** (continua): promedio del saldo anual en su cuenta corriente (CLP).
- **Tiene Crédito Hipotecario** (binaria): reconoce si el cliente tiene tomado un crédito hipotecario con el banco.
- **Tiene Crédito de Consumo** (binaria): reconoce si el cliente tiene tomado un crédito de consumo con el banco.
- **Medio de Contacto Preferente** (categórica): indica el medio preferente del cliente para ser contactado por el banco, las opciones son por celular, e-mail y teléfono particular.
- **Contactos con su Ejecutivo** (continua): número de contactos que realizó el cliente con su ejecutivo en el último año.
- **Tiene Inversiones** (binaria): reconoce si el cliente tiene depósitos a plazo con el banco.

I.2. Transformación y creación de variables

Luego, se realizan algunas transformaciones a las variables originales:

- A partir de la fecha de nacimiento, se calcula la edad que tendría cada cliente al 1-enero-2010 (dado que se indica que los datos fueron creados en ese año).
- Para las variables binarias se pasa de valores en texto (no/si) a dummies 0/1.
- Se crean dummies para cada uno de los posibles valores de las 4 variables categóricas.
- Luego, para las variables de Actividad Laboral y Nivel Educacional se crean variaciones.
  - En el caso de la actividad laboral, se pasa de 7 dummies a sólo 4, agrupando en una sola categoría empresarios y gerentes, y en otra a obreros, técnicos y trabajadores independientes (jubilados y trabajadores dependientes siguen siendo una dummy cada uno).
  - En el caso del nivel educacional se incluyen en una sola variable aquellas personas que cuentan con educación básica o media, pasando de 4 dummies a sólo 3.
  - Por último se crea una tercera variación del nivel educacional con una variable ordinal que ordena los cuatro niveles de educación originales (básica, media, técnica profesional, universitaria).

I.3. Sets de variables.

Una vez que se han realizado las transformaciones mencionadas, se crean tres sets de variables:

- Set 1. Edad, 7 dummies de actividad laboral, 3 dummies de estado civil, 4 dummies de nivel educacional, saldo medio anual, contactos con ejecutivo y las dummies de mora, crédito hipotecario, crédito de consumo e inversiones.
- Set 2. Edad, **4 dummies de actividad laboral**, 3 dummies de estado civil, 4 dummies de nivel educacional, saldo medio anual, contactos con ejecutivo y las dummies de mora, crédito hipotecario, crédito de consumo e inversiones.
- Set 3. Edad, 4 dummies de actividad laboral, 3 dummies de estado civil, **variable ordinal de nivel educacional**, saldo medio anual, contactos con ejecutivo y las dummies de mora, crédito hipotecario, crédito de consumo e inversiones.

Cabe mencionar además que las variables en todos los casos fueron normalizadas. Además para el caso de las variables categóricas transformadas a dummies, se excluye el primer valor, el cual será tomado como valor de referencia. Por ejemplo con la variable Estado Civil se construyeron las dummies Casado, Divorciado y Soltero, sin embargo se mantienen solo las últimas dos.

Por último, las variables binarias creadas a partir de las categóricas son divididas por el la cantidad de categorías posibles con el objetivo de no alterar los resultados producto de la cantidad de categorías (por ejemplo para el caso del Estado Civil, las dummies son divididas por 3).

II. Estimaciones

Es importante tener en cuenta que el método de *K means* consiste en asignar k observaciones de manera pseudo-aleatoria y asignarlas como centros iniciales para el algoritmo. Luego, las observaciones más cercanas a cada centro se asocian al cluster indicado por el centro. Iterativamente, se actualizan los centros de cada cluster para eventualmente converger al centro definitivo de cada segmento.

II.1. Elección del k y del set de variables

Dada la importancia del valor de k, se recurre al método del codo para la elección de este. Esto considera un análisis de la suma cuadrática de las distancias para distintos valores de k.

Para este método se utiliza como métrica de dispersión de los clusters, la suma cuadrática de las distancias de cada observación al centro de su cluster. Tomando en consideración que esta suma **debe** disminuir a medida que aumenta el K, se busca un punto de inflexión K^* tal que la SSD se mantenga "estable" desde $K^* + 1$.

$$\mathrm{SSD}_i = \sum_{j \in \mathcal{C}_i} (x_j - \overline{x}_i)^2$$

A continuación se presentan los resultados del codo para los tres sets considerados.

Figura 1. Análisis del codo (set 1).

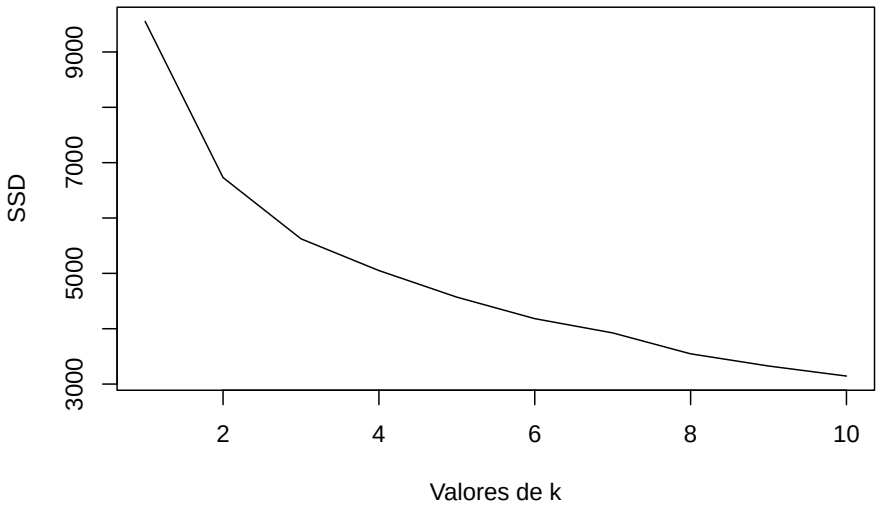


Figura 2. Análisis del codo (set 2).

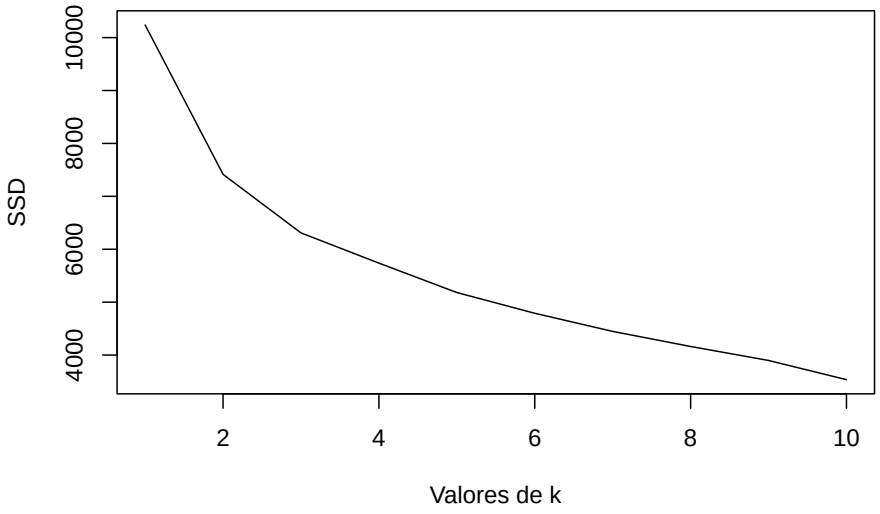
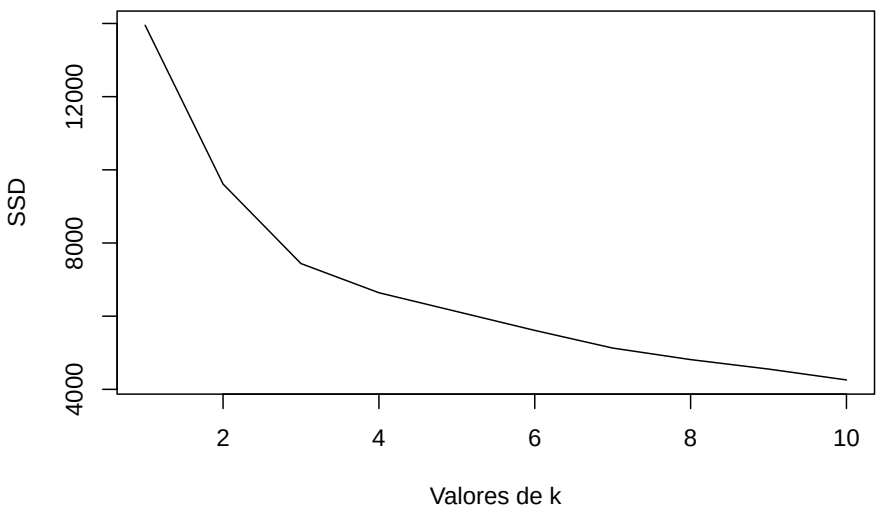


Figura 3. Análisis del codo (set 3).



Para los tres sets de datos considerados se observa un quiebre de la SSD para un k=3 (es decir el codo). A continuación se analiza el valor de la SSD con un k=3 en cada set.

- Set 1. Valor en k=3, la SSD es igual a 5624.6
- Set 2. Valor en k=3, la SSD es igual a 6308.96
- Set 3. Valor en k=3, la SSD es igual a 7437.85

Ante estos resultados se considera que el primer set permite una mayor precisión en la definición de los segmentos.

II.2. Estimación del modelo elegido y análisis de los resultados

Luego de la elección del k=3 y del set 1 de variables, se estiman los clusters. Con esto se obtienen tres clústeres que agrupan a las observaciones analizadas. Los resultados se estudian a continuación.

Tamaño de los clústeres.

Se observa que el cluster de mayor tamaño acumula un 45,9% de los clientes, el siguiente un 38% y el de menor tamaño un 16%.

Table 1: Tamaño de cada cluster.

Cluster	N
1	7245
2	20764
3	17197

Cohesión y separación de los clústeres.

El valor total de las sumas cuadráticas de las distancias con tres clústeres es de **5624.6**. A modo de referencia, la suma cuadrática de distancias inicial era de **9552.37** (es decir la suma cuadrática de las distancias si se tuviera un sólo gran cluster). A nivel de cada cluster, las sumas cuadráticas se presentan a continuación, siendo el cluster de mayor tamaño el que acumula el mayor valor.

Esta disminución de un **41.1%** se puede observar en la Figura 1 incluída en el análisis del codo para la elección del K.

Table 2: Suma cuadrática de las distancias, por cluster.

Cluster	Distancias
1	1309.461
2	2241.696
3	2073.440

Clusters obtenidos

A partir de los datos se observa que las variables más determinantes en la construcción de los clusters son las dummies de crédito hipotecario y de crédito de consumo así como también la variable que indica el saldo medio anual. Dado que el cluster número 1 es el grupo más pequeño (con el 16% de los clientes), se escogerá este segmento para la elaboración de una estrategia comercial.

Según la Tabla 3, para el Cluster Número 1:

- Todos los clientes tienen un crédito de consumo con el banco (ningún otro cluster tiene clientes con crédito de consumo).
- Un 60% de los clientes cuentan además con un crédito hipotecario con el banco (esto ocurre con el 100% de un cluster y el 0% de otro).
- Es el segmento que acumula la mayor cantidad de clientes con educación básica o media (29% vs. el 24% y el 19% de los otros grupos).
- Es el segmento con menor proporción de clientes con educación universitaria (57% vs. 66% en los otros dos segmentos).
- Es el cluster con mayor mora (11% vs. el 5% y el 2% de los otros dos segmentos).
- Es el grupo con menor saldo medio anual ($774.345 vs. $1.258.151 y $1.738.866 en los otros grupos).

Ante esta información, se define el cluster número 1 como el segmento de bajos ingresos.

Table 3: Centros de cada cluster.

	1	2	3
edad	40.19	39.87	41.74
act_laboral_empre	0.04	0.03	0.02
act_laboral_geren	0.09	0.12	0.16
act_laboral_jubi	0.01	0.01	0.04
act_laboral_obre	0.11	0.10	0.04
act_laboral_tecni	0.12	0.09	0.10
act_laboral_tdepen	0.60	0.63	0.59
act_laboral_tinde	0.03	0.03	0.04
e_civil_casa	0.64	0.59	0.59
e_civil_divor	0.13	0.12	0.11
e_civil_solte	0.23	0.29	0.30
n_educ_bas	0.08	0.08	0.09
n_educ_med	0.21	0.16	0.10
n_educ_tp	0.14	0.11	0.15
n_educ_uni	0.57	0.66	0.66
mora	0.11	0.05	0.02
s_medio_anual	774344.79	1258150.50	1738866.43
c_hipote	0.60	1.00	0.00
c_consumo	1.00	0.00	0.00
contactos_eje	0.52	0.66	0.51
inversiones	0.03	0.05	0.13

III. Estrategia de marketing para el Segmento de Bajos Ingresos

- Considerar que esto depende del contexto del negocio y del perfil detectado Debe incluir

III.1. Productos o servicios a ofrecer

Considerando los hallazgos del grupo en el apartado anterior:

- Grupo de bajos ingresos
- Con un crédito de consumo en el 100% de los casos
- Con más de la mitad con créditos hipotecarios
- Con un nivel de mora mayor al resto de los grupos (aunque no mayoritario)
- Menor nivel educacional

Se estima que este grupo es contempla clientes frecuentes del banco, por lo que es importante que tengan una relación sana con su deuda en el banco.

Dado esto se propone un estudio posterior que busque predecir con precisión cuando un cliente está en riesgo de caer en mora y con ello anticipar esta situación. Eventualmente esto podría decantar en un plan de educación financiera para aquellos clientes que se consideren como riesgosos, además de un protocolo más estricto en aquellos créditos de consumo de altos montos para clientes identificados como riesgosos.

III.2. Canales de contacto

Se observa que para este cluster, 70.1% de los clientes han indicado el celular como el medio de contacto preferente (ver Tabla 4). Por lo tanto el canal de contacto preferente será por esta vía y en segundo lugar vía correo electrónico.

Table 4: Medio de contacto para el cluster de bajos ingresos.

m_contacto	n
Celular	5079
Emailing	1726
Fono Particular	440

III.3. Promociones

Además del acompañamiento propuesto en la sección III.1. se sugiere incentivar a aquellos clientes que no están en situación de mora (cerca del 90%) a que mantengan sus cuentas al día con un sistema de alertas e información vía celular (pudiendo ser por medio de una aplicación o SMS). Aquellos clientes que cumplan con sus compromisos bancarios pueden ser premiados con algún sistema de puntos además de obtener mayores facilidades para el acceso a próximos créditos.