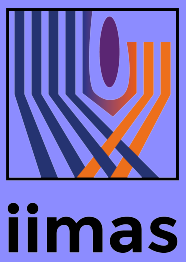
Universidad Nacional Autónoma de México

Práctica 4
Formato empresarial

Aprendizaje de maquinas



Acosta Imandt Daniel
Avitua Varela Fernando
Andres Urbano Guillermo
Barón Bárcenas Iván
Garduño Martinez Eduardo

24 de noviembre de 2022

El siguiente reporte viene de el análisis que se hizo en esta liga

Motivación

Estamos interesados en crear herramientas para saber si los clientes de un banco en especifico son buenos clientes o malos, para eso contamos con una gran cnatidad de datos sobre las transaciiones que han hecho a lo largo del tiempo, al msimo tiempo contamos con infromación del usuario como su género, edad etc.
Para tratar de predecir si son buenos o malos clientes vamos a utilizlar dos métodos diferentes no supervisados, para poder decirle al banco si consideramos que son buenos o malos cleitnes y así que el banco pueda decidir que hacer con est ainformación.

Diccionario de datos

• account: Cada registro describe las características estáticas de una cuenta.
– account id: Identificador de la cuenta.
– district id: Ubicación de la sucursal.
– date: Fecha de creación de la cuenta (en formato YYMMDD).
– frequency: Frecuencia de emisión de estados de cuenta (”POPLATEK MESICNE”
- Emisi on mensual; ”POPLATEK TYDNE” - Emisi on semanal; ”POPLATEK PO OBRATU”:
emisi on despu es de la transacción). • client: Cada registro describe las características de un cliente.
– client id: Identificador del cliente.
– district id: Dirección del cliente.
– birth number: Fecha de nacimiento del cliente.

– gender: Genero (M para masculino, F para femenino).
• disp: Cada registro relaciona a un cliente con una cuenta, es decir, esta relación describe los derechos de los clientes para operar cuentas.
– disp id: Identificador del registro.
– client id: Identificador del cliente.
– account id: Identificador de la cuenta.
– type: Tipo de disposición (propietario / usuario). Solo el propietario puede emitir pedidos permanentes y solicitar un préstamo.
• order: Cada registro describe las características de una orden de pago.
– order id: Identificador del registro.
– account id: Cuenta para la que se emite el pedido.
– bank to: Banco del destinatario. Cada banco tiene un código único de dos letras.
– account to Cuenta del destinatario.
– amount: Cantidad debitada de la cuenta del pedido.
– K symbol: Caracterización del pago. 'POJISTNE' significa Pago de seguro; 'SIPO' son las siglas de pago doméstico; 'LEASING' significa Pago de arrendamiento; 'UVER' significa Pago de Préstamo.

• trans: Cada registro describe una transacción en una cuenta.
– trans id: Es el identificador del registro.
– account id: Cuenta en la que se emite la transacción.
– date: Fecha de la transacción en el formato: AAMMDD.
– type: Tipo de transacción. 'PRIJEM' significa Crédito; 'VYDAJ' significa Débito (retiro).
– operation: Modo de transacción. 'VYBER KARTOU' significa Retiro de tarjeta de crédito; 'VKLAD' son las siglas de crédito en efectivo; 'PREVOD Z UCTU' son las siglas de cobro de otro banco; 'VYBER' significa Retiro en efectivo; 'PREVOD NA UCET' significa remesa a otro banco.
– amount: Monto de la transacción. – balance: Saldo de la cuenta después de la transacción.
– K Symbol: Caracterización de la transacción.'POJISTNE' significa Pago de seguro; 'SLUZBY' significa Pago de Declaración; 'UROK' significa interés acreditado; 'SANKC. UROK significa interés de sanción si saldo negativo; 'SIPO' son las siglas de pago doméstico; 'DUCHOD' significa Pago de pensión de vejez; 'UVER' significa Pago de Préstamo. – bank: Banco del socio. Cada banco tiene un código único de dos letras.
– account: Cuenta del socio.
• loan: Cada registro describe un préstamo otorgado para una cuenta determinada.
– disp id: Identificador del registro.
– loan id: Identificador del crédito.
– account id: Identificador de la cuenta.
– date: Fecha en la que el crédito fue otorgado en formato YYMMDD.
– amount: Monto del crédito.
– duration: Duración del crédito.
– payments: Pagos mensuales del préstamo.
– status: Estado en la liquidación del préstamo. 'A' significa contrato terminado, sin problemas; 'B' significa contrato terminado;, préstamo no pagado; 'C' significa contrato en ejecución, OK hasta ahora; 'D' significa contrato en ejecución, cliente endeudado.
• card: Cada registro describe una tarjeta de crédito emitida a una cuenta.
– card id: Identificador de la tarjeta.
– disp id: Disposición a una cuenta.

– type: Tipo de tarjeta. Los tipos son 'Junior', 'Classic' y 'Gold'.
– emisión: Fecha de emisión de la tarjeta en el formato AAMMDD.
– disp id: Disposición a una cuenta.
• district: Cada registro describe las características demográficas de un distrito..
– A1 = district id: Identificador de distrito
– A2: Nombre del distrito
– A3: Región
– A4: No. de habitantes
– A5: No de municipios con habitantes ¡499
– A6: No. de municipios con habitantes 500-1999

– A7: No de municipios con habitantes 2000-9999
– A8: No de municipios con habitantes¿ 10000
– A9: No. de ciudades
– A10: Ratio de habitantes urbanos
– A11: Salario promedio
– A12: Tasa de desempleo en 1995
– A13: Tasa de desempleo en 1996
– A14: No de emprendedores por 1000 habitantes
– A15: Número de delitos cometidos en 1995
– A16: Número de delitos cometidos en 1996

Solución

Como se puede ver tenemos una gran cantidad de datos de muchas fuentes diferentes por lo que lo primero que tenemos que hacer es ver la información que nos sirve de las diferentes tablas y la que no para nuestro modelo, una vez que tenemos las variables que nos interesan es momento de juntar las tablas para que los datos sean del cliente correcto.La cual la forma correcta de juntarlas es la siguiente:

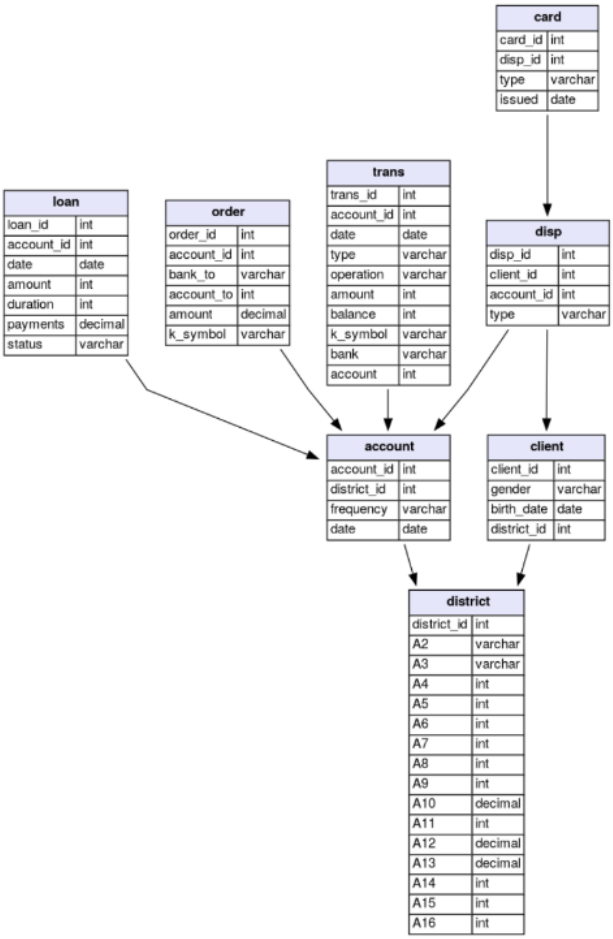


Figura 1: Base de datos relacional

Una vez que se ha hecho lo que se dijo anteriormente quedo la siguiente información para cada cliente:

	account_id	district_id	frequency	loan_id	payments	status	trans_id	type	operation	amount_x	balance	card_id	disp_id	type_x	client_id	type_y	order_id	amount_y	k_symbol	A1	A11
0	5891	54	POPLATEK MESICNE	6202	5,432.00	A	1736607	PRIJEM	VKLAD	900.00	900.00	874	7127	gold	7127	OWNER	38118	5,432.30	UVER	54	9897
1	5891	54	POPLATEK MESICNE	6202	5,432.00	A	1736609	PRIJEM	PREVOD Z UCTU	32,594.00	33,494.00	874	7127	gold	7127	OWNER	38118	5,432.30	UVER	54	9897
2	5891	54	POPLATEK MESICNE	6202	5,432.00	A	1736950	VYDAJ	VYBER	4,500.00	28,994.00	874	7127	gold	7127	OWNER	38118	5,432.30	UVER	54	9897
3	5891	54	POPLATEK MESICNE	6202	5,432.00	A	3673340	PRIJEM	VYBER	21.60	29,015.60	874	7127	gold	7127	OWNER	38118	5,432.30	UVER	54	9897
4	5891	54	POPLATEK MESICNE	6202	5,432.00	A	1736610	PRIJEM	PREVOD Z UCTU	32,594.00	61,609.60	874	7127	gold	7127	OWNER	38118	5,432.30	UVER	54	9897
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
104899	10243	44	POPLATEK MESICNE	7091	3,151.00	C	3088187	PRIJEM	VKLAD	33,098.00	119,158.40	1165	12291	classic	12599	OWNER	44578	3,151.30	UVER	44	8254
104900	10243	44	POPLATEK MESICNE	7091	3,151.00	C	3088310	VYDAJ	VYBER	34,800.00	84,358.40	1165	12291	classic	12599	OWNER	44578	3,151.30	UVER	44	8254
104901	10243	44	POPLATEK MESICNE	7091	3,151.00	C	3088311	VYDAJ	VYBER	34,900.00	49,458.40	1165	12291	classic	12599	OWNER	44578	3,151.30	UVER	44	8254

Figura 2: Una muestra de la información por cliente

Como se puede ver se pudo eliminar gran parte de la información que creíamos que era redundante o no aportaba mucho para el entendimiento de los clientes y poder hacer predicciones de si son bueno clientes o malos.
Para poder trabajar ahora vamos a requerir poder poner todas las variables de la misma forma,

para eso vamos a dividirlas de dos formas las numéricas y las categóricas.
Ya cuando se tiene eso vamos a sacar el promedio de todas sus transacciones(por ejemplo sacar el pago promedio de todas los depósitos que han hecho), para las categóricas vamos a utilizar la moda de todas las transacciones o opciones que hay(por ejemplo ver el tipo de operación que más hace), volvemos a juntar los dos tipos de variables en una sola tabla para los 170 clientes que se nos fueron asignados.

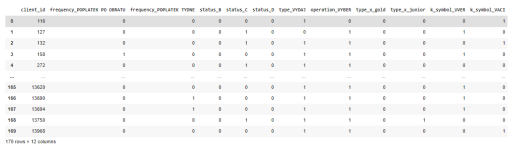


Figura 3: Tabla de variables categóricas

	client_id	payments	amount_x	balance	amount_y	A1	A11
0	116	8,573.00	1,436.00	40,773.30	1,436.00	74.00	10,673.00
1	127	7,348.00	4,700.00	24,915.00	7,348.00	21.00	9,104.00
2	132	4,516.00	3,050.00	45,009.00	3,050.00	36.00	9,198.00
3	158	7,370.00	7,370.20	48,969.30	7,370.20	40.00	9,317.00
4	272	9,112.00	6,500.00	49,345.50	9,020.00	70.00	10,177.00
...	...	...	...	...	...	...	...
165	13620	8,192.00	8,191.50	72,823.60	8,191.50	16.00	8,427.00
166	13690	3,745.00	10,097.00	70,915.70	3,744.70	70.00	10,177.00
167	13694	3,745.00	5,157.50	59,824.95	3,744.70	1.00	12,541.00
168	13750	6,541.00	6,541.20	64,097.40	8,175.60	12.00	8,754.00
169	13968	4,502.00	4,502.30	31,987.60	4,502.30	61.00	8,814.00

170 rows × 7 columns

Figura 4: Tabla de variables numéricas

	client_id	frequency_POPLATEK PO OBRATU	frequency_POPLATEK TYDNE	status_B	status_C	status_D	type_VYDAJ	operation_VYBER	type_x_gold	type_x_junior	k_symbol_UVER	k_symbol_VACI	payments	amount_x	balance	amount_y	A1	A11
0	116	0	0	0	0	0	1	0	0	0	0	1	8,573.00	1,436.00	40,773.30	1,436.00	74.00	10,673.00
1	127	0	0	0	1	0	0	1	0	0	1	0	7,348.00	4,700.00	24,915.00	7,348.00	21.00	9,104.00
2	132	0	0	0	1	0	1	1	0	0	0	1	4,516.00	3,050.00	45,009.00	3,050.00	36.00	9,198.00
3	158	1	0	0	0	0	1	1	0	0	1	0	7,370.00	7,370.20	48,969.30	7,370.20	40.00	9,317.00
4	272	0	0	0	1	0	1	1	0	0	0	0	9,112.00	6,500.00	49,345.50	9,020.00	70.00	10,177.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
165	13620	0	0	0	0	0	1	1	0	0	1	0	8,192.00	8,191.50	72,823.60	8,191.50	16.00	8,427.00
166	13690	0	1	0	0	0	1	1	0	0	1	0	3,745.00	10,097.00	70,915.70	3,744.70	70.00	10,177.00
167	13694	0	1	0	0	0	1	1	0	0	1	0	3,745.00	5,157.50	59,824.95	3,744.70	1.00	12,541.00
168	13750	0	0	0	1	0	1	1	0	1	0	0	6,541.00	6,541.20	64,097.40	8,175.60	12.00	8,754.00
169	13968	0	0	0	0	0	1	1	0	0	0	1	4,502.00	4,502.30	31,987.60	4,502.30	61.00	8,814.00

170 rows × 18 columns

Figura 5: Tabla completa

Por ultimo para la transformación de los datos vamos a estandarizar todas las variables, menos el id del usuario.
Ahora que ya se tienen los datos de la manera más limpia posible y como nos interesa tenerlos, ahora si podemos empezar a crear modelos no supervisados para poder segmentar a los clientes como buenos o malos.

Modelos predictivos

K-medias

Para el primer modelo vamos a utilizar el algoritmo de Kmedias,en pocas palabras lo que se hace es darle en cuantos segmentos queremos que separe un conjunto de puntos (en este caso los clientes) y asignare un segmento a cada punto y juntarlos por aquellos que se encuentren más cerca entre si o compartan características parecidas.
Para este caso en particular quisimos poner cuatro opciones diferentes, ya que los estatus de las compras eran cuatro diferentes, aquellos que se termino la transacción sin ningún problema,

los que se termino con adeudo, aquellos que todavía no termina pero van bien y los que en caso contrario todavía no termina y ya hay ciertos adeudos.Con este modelo se obtuvo los resultados.

```
array([2, 0, 2, 1, 1, 0, 0, 2, 0, 2, 0, 2, 3, 0, 1, 1, 0, 0, 0, 1, 0, 2,
       0, 1, 0, 2, 2, 1, 2, 1, 2, 2, 2, 1, 0, 1, 1, 1, 2, 2, 2, 0, 2, 1,
       2, 2, 0, 2, 2, 0, 2, 2, 1, 2, 0, 1, 1, 1, 2, 1, 1, 0, 1, 0, 0, 1,
       0, 0, 1, 2, 1, 1, 2, 1, 2, 0, 1, 1, 2, 3, 1, 1, 0, 2, 1, 0, 2, 0,
       1, 0, 1, 1, 0, 2, 0, 0, 0, 0, 1, 0, 2, 0, 1, 1, 1, 0, 1, 1, 0, 1,
       0, 0, 1, 1, 2, 2, 2, 0, 0, 2, 1, 1, 0, 1, 1, 2, 0, 0, 0, 0, 2, 2,
       0, 2, 0, 2, 2, 2, 0, 2, 1, 2, 2, 2, 1, 0, 1, 1, 1, 1, 2, 2, 1, 0,
       1, 1, 1, 1, 2, 1, 2, 2, 0, 2, 1, 1, 1, 1, 1, 2], dtype=int32)
```

Figura 6: Los clústers de los usuarios

Agrupamiento jerárquico

Para el siguiente modelos decidimos utilizar *Agrupamiento jerárquico* este va juntando puntos que se encuentren más cercanos en el espacio poco a poco, hasta que solo queden cuatro grupos diferentes, por lo que queda de la siguiente manera los grupos con este algoritmo.

```
array([2, 3, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 3, 1, 1, 0, 2,
       0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1,
       1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0,
       1, 0, 1, 0, 0, 1, 0, 0, 1, 3, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1,
       1, 0, 1, 0, 0, 0, 1, 0, 3, 0, 0, 0, 0, 0, 1, 1, 0, 3, 1, 0, 0, 1,
       0, 0, 1, 1, 2, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 3, 3, 0, 1, 0,
       0, 0, 3, 0, 0, 0, 0, 0, 3, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 2, 1, 0,
       1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0])
```

Figura 7: Los clústers de los usuarios

Juntar los dos modelos

Por ultimo vamos ajuntar los dos algoritmos para obtener un mejor resultado, es decir vamos a sacar kmedias a los resultados que obtuvimos con el segundo método, de esta forma se obtuvo lo siguiente

```
array([1, 0, 1, 2, 2, 0, 0, 1, 0, 1, 0, 1, 0, 0, 2, 2, 0, 0, 0, 2, 0, 1,
       0, 2, 0, 1, 1, 2, 1, 2, 1, 1, 1, 2, 0, 2, 2, 2, 1, 1, 1, 0, 1, 2,
       1, 1, 0, 1, 1, 0, 1, 1, 3, 1, 0, 3, 2, 2, 1, 2, 2, 0, 3, 0, 0, 3,
       0, 0, 2, 1, 3, 2, 1, 3, 1, 0, 2, 2, 1, 0, 2, 3, 0, 1, 2, 0, 1, 0,
       2, 0, 2, 3, 0, 1, 0, 0, 0, 0, 3, 0, 1, 0, 2, 2, 3, 0, 2, 3, 0, 2,
       0, 0, 2, 2, 1, 1, 1, 0, 0, 1, 3, 2, 0, 3, 3, 1, 0, 0, 0, 0, 1, 1,
       0, 1, 0, 1, 1, 1, 0, 1, 2, 1, 1, 1, 2, 0, 2, 2, 2, 2, 1, 1, 2, 0,
       2, 2, 3, 2, 1, 3, 1, 1, 0, 1, 2, 2, 2, 2, 3, 1], dtype=int32)
```

Figura 8: Los clústers de los usuarios

Conclusiones

Por ultimo ya tenemos como dividiríamos a los diferentes usuarios, y queda de la siguiente manera.

client_id	frequency_POPLATEK PO OBRATU	frequency_POPLATEK TYDNE	status_B	status_C	status_D	type_VYDAJ	operation_VYBER	type_x_gold	type_x_junior	k_symbol_UVER	k_symbol_VACI	payments	amount_x	balance	amount_y	A1	A11	clusters
116	0	0	0	0	0	1	0	0	0	0	1	8,573.00	1,436.00	40,773.30	1,436.00	74.00	10,673.00	1
127	0	0	0	1	0	0	1	0	0	1	0	7,348.00	4,700.00	24,915.00	7,348.00	21.00	9,104.00	0
132	0	0	0	1	0	1	1	0	0	0	1	4,516.00	3,050.00	45,009.00	3,050.00	36.00	9,198.00	1
158	1	0	0	0	0	1	1	0	0	1	0	7,370.00	7,370.20	48,969.30	7,370.20	40.00	9,317.00	2
272	0	0	0	1	0	1	1	0	0	0	0	9,112.00	6,500.00	49,345.50	9,020.00	70.00	10,177.00	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
13620	0	0	0	0	0	1	1	0	0	1	0	8,192.00	8,191.50	72,823.60	8,191.50	16.00	8,427.00	2
13690	0	1	0	0	0	1	1	0	0	1	0	3,745.00	10,097.00	70,915.70	3,744.70	70.00	10,177.00	2
13694	0	1	0	0	0	1	1	0	0	1	0	3,745.00	5,157.50	59,824.95	3,744.70	1.00	12,541.00	2
13750	0	0	0	1	0	1	1	0	1	0	0	6,541.00	6,541.20	64,097.40	8,175.60	12.00	8,754.00	3
13968	0	0	0	0	0	1	1	0	0	0	1	4,502.00	4,502.30	31,987.60	4,502.30	61.00	8,814.00	1

ows × 19 columns

Figura 9: Los usuarios con el clúster que le asignamos

Ahora tenemos que ver cual cluster diremos que es el de los usuarios positivos y el de los negativos, para eso vamos a analizar algunas de las estadísticas de cada cluster, con la siguiente tabla.

	payments					amount_x					balance					amount_y				
	min	max	median	mean	std	min	max	median	mean	std	min	max	median	mean	std	min	max	median	mean	std
clusters																				
0	312.000000	7348.000000	2820.000000	2897.490566	1497.286528	373.700000	8300.000000	3697.800000	3518.471698	1656.878142	24056.850000	70605.800000	49663.400000	48489.664151	9750.695065	312.000000	7348.000000	3626.000000	3566.827358	1534.639532
1	501.000000	5710.000000	4509.000000	4348.722222	2201.948583	607.000000	7291.300000	3651.000000	3561.042593	1735.017543	26388.800000	70943.400000	49452.725000	49032.619444	9323.806338	226.000000	7291.300000	3136.000000	3246.062037	1863.135378
2	3151.000000	9910.000000	6258.500000	6204.826087	1481.699989	3300.000000	11021.000000	6929.900000	7146.256522	1588.570600	39825.200000	80277.900000	59661.625000	59776.726087	9730.746657	3151.300000	9794.500000	6682.850000	6717.039130	1571.578419
3	2254.000000	8253.000000	6239.000000	5636.647059	1947.392903	4875.000000	12250.000000	6695.000000	7537.176471	2046.974855	44720.900000	74791.750000	62305.000000	60168.185294	7942.204542	3990.500000	9275.850000	6366.250000	6598.029412	1648.169560

Figura 10: Estadísticas de los clusters

Como podemos observar se logran notar tendencias diferente en los diferentes clusters,lo que implica que si los segmento por diferentes comportamientos del usuario.
En donde diremos que los clusters 0 y 1 son negativos y para los 2 y 3 son buenos clientes.