Proyecto: Uso del internet para solicitar/comprar bienes y servicios

Aprendizaje Automatizado, 2019-II

Erick García Ramírez

MCIC–IIMAS, UNAM

13 de Junio de 2019

Aspectos generales

- Un estudio sobre comercio en línea.
- Base de datos con perfiles socio-económicos de individuos e información sobre su actividad de compras en internet.
- **Objetivos generales**: conocer sobre el perfil de compradores y generar un modelo predictivo.

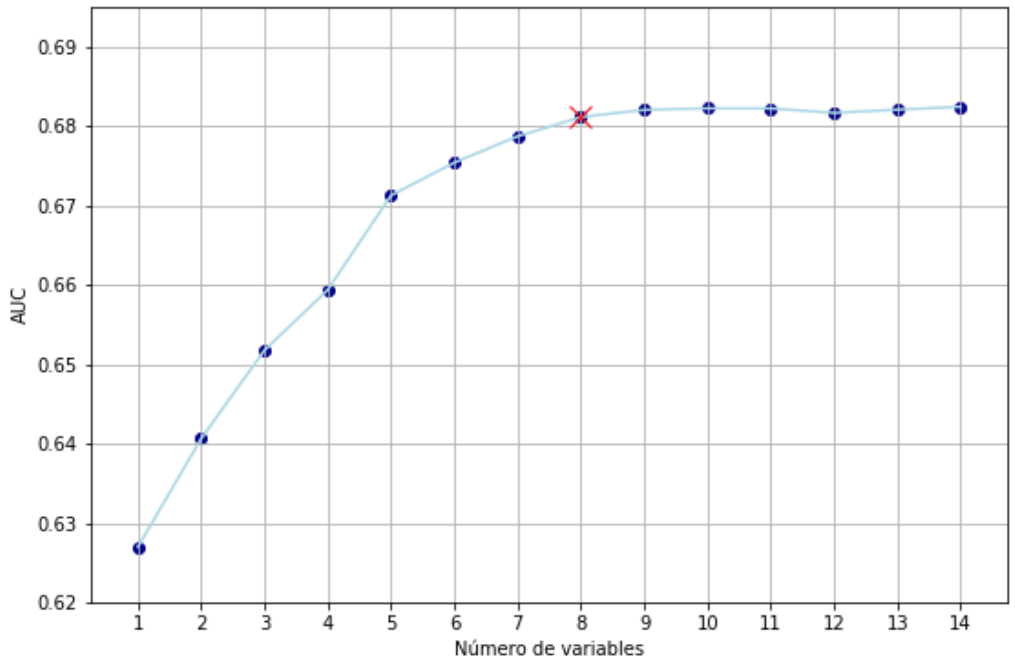
La base de datos

- *Individual Company Sales: disponible en Kaggle.com*
- 14 atributos de clientes, una bandera: 'Y' si compró el producto y 'N' en caso contrario
- 13558 renglones

	flag	gender	education	house_val	age	online	customer_psy	marriage	child	occupation	mortgage	house_owner	region	car_prob	fam_income
5	Y	F	3. Bach	248694	6upto65	Y	B	Married	N	Professional	2Med	Owner	West	1	G
7	N	F	3. Bach	416925	5upto55	Y	C	Married	Y	Professional	1Low	Owner	South	2	I
11	Y	F	3. Bach	245686	4upto45	N	F	Married	U	Blue Collar	1Low	Owner	South	3	E
12	Y	F	2. Some College	360587	5upto55	Y	C	Married	Y	Professional	3High	Owner	Midwest	1	J
15	Y	M	0. lessHS	162884	1_Unk	Y	G	Married	Y	Professional	1Low	Renter	South	7	C

(a) forward-selection y regresión logística

- Para cada $n \in \{1, \ldots, 14\}$, hallar (acumulativamente) los mejores n atributos
- Mejores n atributos si generan el modelo de **regresión logística** con máxima **AUC**
- No checa todos los posibles $2^{14} - 1$ subconjuntos de atributos (pero la complejidad se mantiene baja)



(a) forward-selection y regresión logística

n	***n* atributos más importantes**	**AUC**
1	customer_psy	0.626
2	age	0.640
3	education	0.651
4	online	0.659
5	gender	0.671
6	car_prob	0.675
7	mortgage	0.678
8	**occupation**	**0.681**
9	child	0.682
10	house_val	0.682
11	house_owner	0.682
12	fam_income	0.681
13	region	0.682
14	*todas*	0.684

Acumulando hacia abajo.

Mejores 8 atributos y modelo predictivo

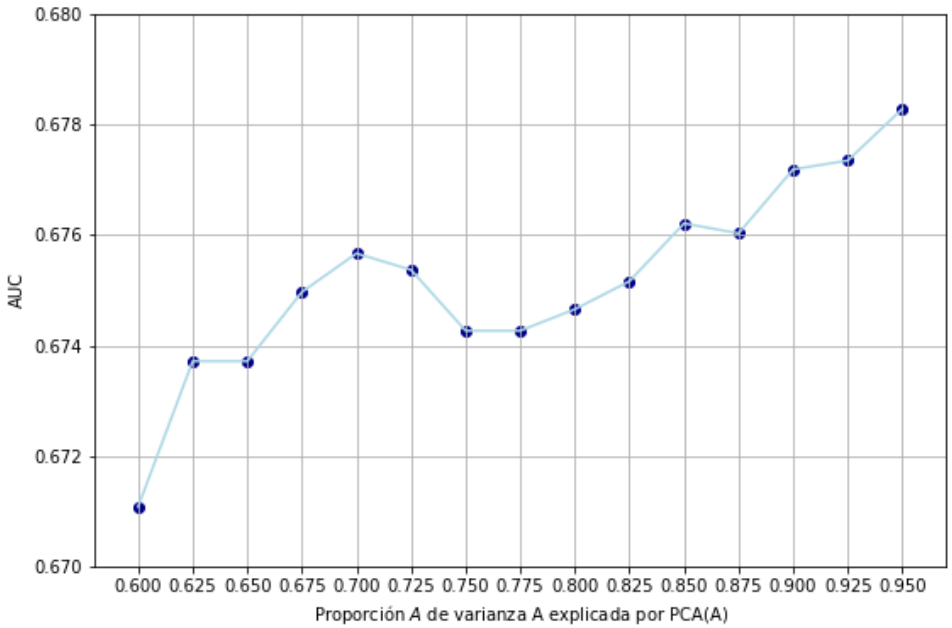
customer_psy, *age*, *education*, *online*, *gender*, *car_prob*, *mortgage*, *occupation*

- El modelo de regresión logística correspondiente alcanza un score de 0.706 sobre el conjunto de prueba.
- La matriz de confusión sobre el conjunto de prueba es:

$$\begin{bmatrix} & Predicted0 & Predicted1 \\ True0 & 1830 & 1275 \\ True1 & 801 & 3162 \end{bmatrix}$$

(b) ¿Se obtiene un mejor modelo usando PCA?

- PCA reduce la dimensionalidad encontrando las componentes (ejes) que observan la mayor varianza entre los atributos.
- Después de codificar todos los atributos categóricos, la matriz de atributos tiene dimensión 42 (columnas).
- Se prueban diferentes valores de la proporción A de varianza explicada. Se busca el modelo con mayor AUC para cada valor de A.



Modelo a partir de PCA

- Se toma mejor modelo con $A = 0{,}90$. Dimensión reducida de 42 a 21.
- El modelo de regresión logística alcanza un score de 0.701 sobre el conjunto de prueba
- la matriz de confusión sobre el conjunto de prueba es:

$$\begin{bmatrix} & Predicted0 & Predicted1 \\ True0 & 1833 & 1272 \\ True1 & 843 & 3120 \end{bmatrix}$$

- Es díficil determinar cuáles atributos reflejan las 21 dimensiones encontradas por PCA.

(c) Segmentación por clustering

- Dividir a los clientes en grupos de acuerdo a sus características
- Objetivo: distinguir clientes por medio de sus rasgos entre compradores y no compradores
- Se usan sólo los 8 atributos más importantes
- Aplicar KModes, una variante de KMeans para datos mixtos.

$$\texttt{kmodes} = \texttt{KModes}(\texttt{n\_clusters} = 2, \texttt{init} =' \texttt{Huang}', \texttt{n\_init} = 5)$$

- Se hallan dos clusters con centros con coordenadas mixtas.
- Es un poco complicado interpretar los clusters. El clustering no es suficientemente bueno para distiguir compradores.

Trabajo futuro

- Muchas variantes se pueden probar para mejorar el modelo predictivo
- Usar best-subset en vez de forward-selection
- Extraer más información del clustering, probar particiones en más clusterings considerando diferentes variables
- Probar otras codificaciones de los atributos categóricos