Proyecto de Aprendizaje Automatizado:
Uso del internet para comprar bienes: perfilado de clientes

Erick García Ramírez*

Resumen

Presentamos una aplicación del Aprendizaje Automatizado y la Minería de Datos al comercio en línea, en particular al problema de determinar el perfil de un cliente. El conjunto de datos considerado contiene la información de los clientes de una empresa, para los cuales deseamos identificar los rasgos más importantes que determinan si comprarán el producto ofertado por la empresa. Para esta tarea empleamos las técnicas de *one-step forward-selection* y regresión logística. Posteiormente aplicamos el algoritmo de clustering KModes para segmentar el conjunto de clientes. Indicamos como un análisis más detallado de los clusters permitiría especificar las características de los clientes que con mayor probabilidad comprarán el producto.

Keywords: *E-commerce, regresión logística, forward-selection, datos mixtos, clustering, KModes*

1. Introducción

A lo largo de las últimas décadas las aplicaciones del Aprendizaje Automatizado han sido bastas y de naturaleza muy diversa. Algunas de ellas, como lo son los Sistemas Automatizados de Recomendación y la Detección de Operaciones Irregulares, tienen el objetivo de apoyar a las actividades comerciales de empresas que ofertan sus bienes y/o servicios por medio del internet. En este trabajo desarrollaremos una aplicación del Aprendizaje Automatizado a este rubro.

Para una empresa que oferta algún producto o un servicio resulta muy valioso distinguir—de entre todas las personas a las que dirige su publicidad—a aquellas personas que con mayor probabilidad comprarán o contrarán su oferta. Determinar las características de los *buenos clientes* y los *malos clientes* es importante pues le da a la empresa la oportunidad de optimizar sus estrategias de publicidad y de administración de sus clientes.

Con la proliferación de grandes datos sobre el consumo de bienes y contratación de servicios, y en particular con la enorme disponibilidad de datos de comercio en línea, las tareas de distinguir las características importantes de clientes es una aplicación clásica y bien desarrollada de la Minería de Datos y el Aprendizaje Automatizado ([1, Cap. I]).

En el presente trabajo aplicaremos técnicas de Aprendizaje Automatizado a los datos de venta de un producto por parte de una empresa, con el objetivo de descubrir las características de las personas que con mayor probabilidad comprarán el producto.

2. Antecedentes

El uso de la Minería de Datos y el Aprendizaje Automatizado para apoyar actividades comerciales tiene un impacto importante y una larga tradición, vea por ejemplo [1][2]. Dentro de este contexto, el *perfilado de usuarios o clientes* es una de las tareas en que dichas disciplinas sobresalen. [4][5][6][7] y [9] exploran la aplicación del aprendizaje automatizado al perfilado de usarios o clientes.

Un *perfil* es un conjunto de información que funge como representación de una persona, usuario o cliente[3]. La información que conforma un perfil puede ser de naturaleza diversa; puede incluir, por

*erick_phy@ciencias.unam.mx, MCIC, IIMAS–UNAM.

ejemplo, datos conductuales, rasgos físicos y/o rasgos socio-económicos. La información que debe constituir un perfil típicamente se determina a partir del uso que se desea dar a dicho perfil. Por ejemplo, un perfil de un candidato a ingresar a alguna universidad incluye información sobre su rendimiento en niveles acádemicos precedentes, su rendimiento en exámenes de ingreso y, posiblemente, algunos datos sobre su contexto socio-económico. En este caso la información necesaria para un perfil es especificada por el cómite encargado del proceso de selección de la universidad.

En muchos otros contextos especificar la información que debe constituir un perfil no es tan sencillo como en el ejemplo del párrrafo anterior.

El siguiente es el ejemplo que abordaremos en este trabajo. Cuando una empresa tiene la intención de lanzar un nuevo producto a la venta, o de simplemente mejorar la estrategia publicitaria sobre algún producto, hay un gran interés en determinar qué personas serán sus potenciales compradores, es decir, es importante descubrir el perfil de los compradores del producto. El descubrimiento exitoso de dicho perfil permitiría a la empresa diseñar una estrategia publicitaria adecuada para maximizar sus ventas del producto. Sin embargo, puede resultar complicado determinar por adelantado y con precisión cuáles serán los rasgos del perfil que más influirán en la decisión de compra del producto. Frente a tal problema, el punto de vista más común en la actualidad es el de colectar la mayor información posible para formar un perfil provisional, para posteriormente determinar un perfil más preciso (en este caso, el perfil de las personas que con mayor probabilidad comprarán el producto).

Entre las técnicas de Minería de Datos y Aprendizaje Automatizado utilizadas para el perfilado de usuarios y/o clientes se encuentran los árboles de decisión, los algoritmos de clustering y los algoritmos de clasificación (vea [6], donde puede encontrar una tabla comprensiva sobre las técnicas usadas para esta tarea). Cada uno de estas tiene cualidades o desventajas que las hacen valiosas o inapropiadas dependiendo de sus requisitos y condiciones generales del problema. Los árboles de decisión tienen la característica de dar una buen panorama de las características del perfil, con resultados interpretables de manera natural y de naturaleza cualitativa [1, Cap. 6]. Los algoritmos de clustering ofrecen una segmentación de los usuarios/clientes, subrayando aquellas características que les hacen ser similares. Los algoritmos de clasificación (e.g. máquinas de soporte vectorial y redes neuronales) al ser combinadas con alguna técnica de selección de variables ofrecen también una forma de determinar los rasgos importantes de un perfil. A menudo, en combinación, estas y otras técnicas ayudan en la tarea de determinar el perfil más útil.

3. Metodología

El conjunto de datos que usaremos en este trabajo es *Individual Company Sales Data*, disponible en www.kaggle.com [8]. Los datos corresponden a las ventas por internet de un producto por parte de una empresa[1]. Cada renglón en el conjunto corresponde a la información de un cliente. Se colectaron 14 datos sobre el cliente y una bandera que señala si el cliente compró el producto o no. Después de remover aquellos renglones en los que hay algún valor no conocido de algún atributo, el conjunto consta de 23558 muestras. La tabla 1 describe la estructura y naturaleza del conjunto de datos.

Ya que la naturaleza de los datos es mixta, es decir, hay tanto datos númericos como categóricos debemos comentar sobre la preparación y el preprocesamiento de ellos. Se tomaron los siguientes pasos de manera consecutiva.

I. Todos aquellos atributos de tipo categórico ordinal se cambiaron directamente por números en $[0, 1]$ (e.g., en fam_income 'A' corresponde a 0.0 y 'L' a 1.0).

II. house_val se re-escaló al intervalo [0,1].

III. Para aquellos atributos categóricos binarios (e.g., 'gender') se cambiaron los valores originales a 1 y 0.

IV. Para aquellos atributos de tipo categórico nominal que no son binarios se aplicó one-hot-encoding (sin embargo, el punto donde esto se hace depende de la técnica de aprendizaje con la que se trabaja)

Se realizaron las siguientes tareas de Aprendizaje Automatizado el sobre conjunto de datos.

[1]El conjunto se presenta anonimizado, no se conoce ni el producto ni el nombre de la empresa

Columna	Atributo	Tipo de dato	Valores
0	**flag**	categórico nominal	Y, N si compró el producto, (56.6 % y 43.4 % resp.)
1	**gender**	categórico nominal	F, M (39 % y 61 % resp.)
2	**education**	categórico ordinal	0.lessHS, 1.HS, 2.SomeCollege, 3.Bach, 4.Grad (8.7 %, 21.4 %, 27.8 %, 25.4 % y 16.6 % resp.)
3	**house_val**	númerico	en [0, 999999]
4	**age**	categórico nominal	1_Unk, 2upto25, 3upto35, 4upto45, 5upto55, 6upto65, 7above65 (13 %, 2.5 %, 10.7 %, 21.1 %, 25.8 %, 16.9 % y 9.8 % resp.)
5	**online**	categórico nominal	Y, N el cliente tiene experiencia en compras por interne(70.1 % y 29.9 % resp.);
6	**customer_psy**	categórico nominal	A–J, psicología del cliente, basaso en área de residencia (3.4 %, 21.9 %, 22.3 %, 5.6 %, 15.9 %, 9.3 %, 9.9 %, 1.6 %, 4.9 % y 5.1 % resp.)
7	**marriage**	categórico nóminal	Y, N (81.8 % y 18.2 % resp.)
8	**children**	categórico nominal	Y, N, U (31 %, 47 % y 21 % resp.)
9	**occupation**	categórico nominal	Professional, Blue Collar, Retired, SalesService, Others, Farm (41.7 %, 16.9 %, 8.6 %, 28.2 %, 3.9 % y 0.7 % resp.)
10	**mortgage**	categórico nominal	1Low, 2Med, 3High (69.9 %, 13.7 % y 16.3 % resp.)
11	**house_owner**	categórico nominal	Owner, Renter (79.7 % y 20.3 % resp.)
12	**region**	categórico nominal	West, South, Midwest, Northeast, Rest (21.8 %, 39 %, 20.7 %, 17.8 % y 0.6 % resp.)
13	**car_prob**	categórico ordinal	1–9; probabilidad de que comprará un auto nuevo (32.1 %, 18.8 %, 13.4 %, 7.1 %, 6.6 %, 5 %, 4.8 %, 6.3 % y 5.8 % resp.)
14	**fam_income**	categórico ordinal	A–L; nivel del ingreso familiar, L es el más alto (5 %, 4.8 %, 5.7 %, 10.5 %, 20.7 %, 17.4 %, 11 %, 6.8 %, 4.6 %, 4.6 %, 4.1 % y 4.6 % resp.)

Tabla 1: La estructura de la base de datos.

(a) Se halla el subconjunto de atributos *más significativos* mediante un modelo de regresión logística y forward-selection [11, Cap. 6].

(b) Se aplica un Análisis de Componentes Principales PCA [10] y posteriormente se ajusta un modelo de regresión logística (seleccionado mediante validación cruzada).

(c) Se explora la segmentación de los clientes por medio de clustering (K-modes) [12].

4. Resultados

(El código del presente trabajo se encuentra en: https://github.com/erickgrm/machine-learning/tree/master/project)

El conjunto de 23558 renglones se particionó en un conjunto de entrenamiento y otro de prueba con la proporción 70–30 %. En (a) y (b) se trabaja con el conjunto de entrenamiento para la selección de variables. El conjunto de prueba se usa únicamente para observar la bondad del mejor modelo encontrado con respecto a predicciones sobre muestras nuevas.

(a) *Forward-Selection* y regresión logística

El método de forward-selection[11] halla iterativamente el mejor modelo considerando desde uno hasta el total de atributos (en nuestro caso, hasta 14 atributos). El método funciona partiendo de algún modelo de clasificación, que en nuestro caso será el de regresión logística. También se requiere fijar el criterio para comparar los diferentes modelos; para nosotros este criterio será el de mayor área bajo la curva ROC (AUC).

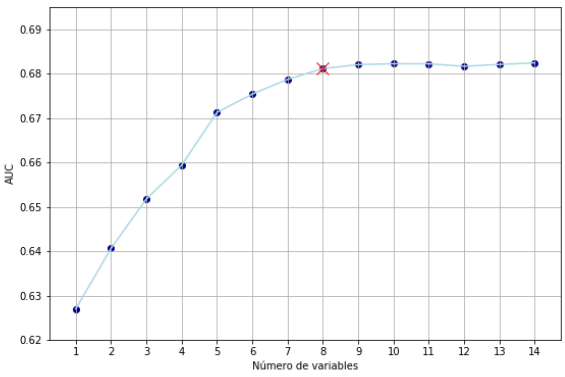

Figura 1: *Área bajo la curva ROC contra número de variables consideradas*

La figura 1 muestra el comportamiento de AUC con respecto al número de variables consideradas. Se puede observar (y analizar numéricamente) que la AUC aumenta significativamente desde 1 hasta 8 atributos, pero su crecimiento se desacelera para más de 8 atributos. La figura 2 soporta esta elección pues muestra que el score de los modelos varia de una forma similar con respecto al número de atributos.

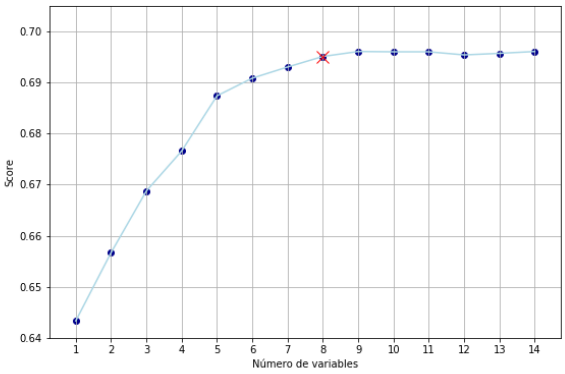

Figura 2: *Score (para regresión logística) contra número de variables consideradas.*

Por lo tanto, tomamos como subconjunto de *atributos más importantes* a los 8 atributos que forward-selection halló considerando 8 atributos. Los atributos son:

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:

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

Adicionalmente, forward-selection da una idea de cuáles son los n atributos más importantes para cada $n \leq 14$, vea la tabla 2.

n	n atr. 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

Tabla 2: *Los n atributos más importantes. La columnas 2 y 3 son acumulativas hacia abajo.*

(b) PCA y regresión logística

El Ánalisis de Componentes Principales (PCA) es una técnica empleada para reducir la dimensionalidad de un conjunto de datos. PCA identifica los ejes en el espacio n-dimensional sobre los cuales el conjunto de datos tiene más variabilidad. Aunque PCA ayuda a eliminar algunos atributos (los que no añaden mucha información, i.e. varianza, al conjunto de datos) es casi siempre díficil interpretar su resultado. No es fácil extraer cuáles fueron los atributos que PCA eliminó. Su valor en nuestro caso, es verificar si después de su aplicación, PCA mejora las cualidades de un modelo de regresión logística.

Despues de aplicar one-hot-encoding al conjunto de datos de entrenamiento se cuenta con 42 columnas de atributos. Se aplica PCA para reducir su dimensionalidad. Aplicamos la versión PCA(A) donde $A \in (0, 1)$ es la cantidad de varianza que deseamos cubrir al reducir la dimensión. Por ejemplo, para $A = 0.85$ se pasa de 42 dimensiones a sólo 18. Para cada valor de A, se genera la matriz (de dimensión reducida) y con ella se encuentra el mejor modelo de regresión logística seleccionado por validación cruzada (*10-fold*). Comparamos los modelos obtenidos por su AUC. La figura 3 muestra el comportamiento

de la AUC de los modelos generados con respecto a A.

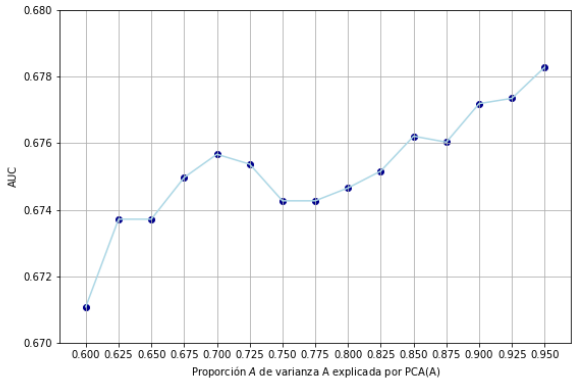

Figura 3: *Área bajo la curva ROC contra proporción de varianza A explicada por PCA*

La máxima AUC se alcanza para $A = 0.95$ y para tal valor la dimensión se redujó de 42 a 26. Para lograr un balance entre maximizar AUC y disminuir la dimensión lo más posible, proponemos tomar como mejor modelo a aquel correspondiente a $A = 0.90$.

Para $A = 0.90$ la dimensión se redujó de 42 a 21, y el AUC alcanzado es de 0.677. Para el modelo de regresión logística correspondiente, el score sobre el conjunto de prueba es de 0.701 y la matriz de confusión sobre el conjunto de prueba es:

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

(c) Segmentación vía Clustering

Aplicamos el algoritmo K-modes, el cual es un algoritmo de clustering basado en K-means pero que está diseñado para poder ser usado con conjuntos de datos mixtos[12]. Usaremos el conjunto de datos limitado al conjunto de atributos más importantes descubierto en (a).

Investigamos el resultado de segmentar a los clientes en dos clusters, buscando capturar los grupos 'flag' = 1 y 'flag= 0' (i.e., separar compradores de no compradores). Se ejecutó Kmodes en cinco ocasiones obteniéndose 126657.0 el menor valor de costo[2]. Dicho valor genera dos clusters identificados por las etiquetas 0 y 1. Para juzgar si esta segmentación refleja de alguna forma la división buscada en compradores y no compradores, podemos comparar el etiquetado del clustering contra la columna 'flag' (n.b. recuerde que el etiquetado en clustering no es siempre el mismo sobre diferentes ejecuciones; en este punto se debe checar cómo se corresponden 1 y 0 del etiquetado del clustering con el 1 y 0 de 1flag').

Los clusters obtenidos por KModes (usando los 8 atributos más importantes) etiquetan correctamente a 14428 de las muestras, mientras que se equivocan en las 9130 restantes. Esto da un rendimiento (score) para KModes de 0.612.

5. Conclusiones

Por medio de forward-selection y regresión logística hallamos que para determinar si un cliente comprará el producto que ofrece la empresa basta considerar los 8 atributos

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

Estos son los atributos que constituyen el perfil de un cliente. Dado un cliente nuevo, la empresa debe colectar los valores de estos atributos para posteriormente hacer una predicción sobre si el cliente comprará el producto o no usando el modelo entrenado para este trabajo. En La tabla 2 también puede orientar la decisión de qué atributos considerar si por alguna razón no se cuenta con todos los 8 mejores.

En (c) se aplicó la técnica de clustering considerando únicamente los 8 atributos más importantes. Un análisis exploratorio de los clusters resultantes, y la prueba con segmentaciones en más de dos clusters, determinaría cuáles valores específicos de los atributos maximizan la probabilidad de que un cliente dado comprará el producto. El resultado de tal análisis podrá entonces ser usado para mejorar la estrategia publicitaria del producto.

En el presente trabajo hemos mostrado un ejemplo de cómo se pueden aplicar algunas técnicas de Aprendizaje Automatizado a un problema de comercio en línea. La extensión está aplicación puede tomar muchas direcciones. En el aspecto técnico, algunas modificaciones al trabajo aquí presentado incluyen:

[2]En KModes la función de costo está determinada como una combinación de la distancia Euclidiana entre atributos números y una medida de similaridad entre atributos categóricos, vea[12, Sec. 2.2].

(i) Usar la técnica *best-subset*[11, Cap. 6] en vez de forward-selection en (a). Esta última no explora todos los $2^{14} - 1$ subconjuntos de atributos posibles. La desventaja es que la complejidad del algoritmo se vuelve entonces exponencial.

(ii) Considerar otros modelos de clasificación para reemplazar a la regresión logística en (a)[3].

(iii) Para (c), aplicar la técnica CENG[13] para codificar numéricamente todos los atributos de tipo categórico. CENG propone hacer esto minimizando la probabilidad de modificar los patrones (covarianzas) originales en el conjunto de datos.

Referencias

[1] M. Berry y G. Linoff. *Data Mining Techniques For Marketing, Sales, and Customer Relationship Management.* 2da Edición. Wiley Publishing Inc., 2004.

[2] J. Dean. *Big Data, Data Mining, and Machine Learning: Value Creation for Business, Leaders and Practitioners.* John Wiley Sons, 2014.

[3] A. Cufoglu. *User Profiling-A Short Review.* International Journal of Computer Applications, Volume 108– No. 3, 2014.

[4] T. Fawcett y P. Foster. *Combining Data Mining and Machine Learning for Effective User Profiling.* En KDD-96 Proceedings, págs. 8-13, AAAI, 1996.

[5] A. Bellogín, I. Cantador, P. Castells y Á. Ortigosa. *Discovering Relevant Preferences in a Personalised Recommender System using Machine Learning Techniques*, en Preference Learning Workshop, en la 8th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2008.

[6] M. Harandi. *User Profiling in News Recommender Systems.* Disponible en: https://www.ntnu.no/wiki/download/attachments/86731314/User%20Profiling%20NRS.pdf?version%3

[7] Q. Chen, A. Norcio y J. Wang. *Neural Network Based Stereotyping for User Profiles.* Neural Computing and Applications No. 9, págs. 259–265, 2000.

[8] *Individual Company Sales Data.* Disponible en: https://www.kaggle.com/mickey1968/individual-company-sales-data.

[9] K. Kashwan y C. Velu. *Customer Segmentation Using Clustering and Data Mining Techniques.* International Journal of Computer Theory and Engineering, Vol. 5 (6), págs. 856–860, 2013.

[10] M. Zaki y W. Meira. *Data minining and Analysis.* Cambridge University Press, 2014.

[11] G. James, D. Witten, T. Hastie y R. Tibshirani. *An Introduction to Statistical Learning.* Springer 2013.

[12] Z. Huang. *Clustering large data sets with mixed numeric and categorical values.* En Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference, págs. 21–34, 1997.

[13] A. Kuri. *Categorical Encoding with Neural Networks and Genetic Algorithms.* Recent Researches in Applied Informatics: Proceedings of AICT '15, WSEAS Press, 2015.

[3] Preliminarmente se consideró máquinas de soporte vectorial, pero su desempeño no fue considerablemente mejor y su entrenamiento es muy lento.