



CURSO : IN720 – Técnicas Avanzadas de Minería de Datos
 PROFESOR : Richard Weber – Fabián Medel
 P. AUXILIARES: Jaime Miranda – Cristián Hormazabal
 SEMESTRE : Primavera 2004

TAREA 1

Una de las áreas más exitosas de aplicación de Datamining (DM) es la de ventas por correspondencia. En esta industria, uno de los aspectos más relevantes son las devoluciones de productos defectuosos o que no cumplieran con las expectativas del cliente, la razón de su importancia son las políticas de servicio orientadas al cliente por parte de las empresas. Sin embargo la administración de estas situaciones involucra altos costos en logística e imagen.

Una compañía de esta industria quiere hacer un pronóstico de las tasas de devolución de sus clientes y con esta información, planificar una estrategia comercial. Para ello lo primero será diferenciar tres segmentos de clientes dentro del total: aquellos con alta tasa de devolución, aquellos con baja tasa y el resto.

Existe un total de 20.146 clientes para los cuales se posee información personal y transaccional (dmc2004_train.txt). Para éstos, se tiene además como atributo la información respecto del número de órdenes entregadas (num_ent) y número de devoluciones (num_dev).

Un primer objetivo es entonces diferenciar dentro de ellos las categorías de interés que se definirán según la regla dada por la razón entre órdenes devueltas y órdenes entregadas (variable ratio), según la siguiente tabla:

Ratio = $\frac{num_dev}{num_ent}$	Tasa de devolución	Clase
= 0.18	BAJA	L (low)
= 0.40	ALTA	H (high)
Resto	INDEFINIDA	U (undefined)

Por lo tanto se debe crear una nueva variable objetivo (Ratio) la cual clasifique a los clientes del conjunto de entrenamiento. Una vez realizado esto, se requiere construir un modelo predictivo utilizando alguna técnica de DM que permita clasificar un segundo conjunto de datos para los cuales no existe la información de entregas y devoluciones realizadas (dmc2004_class.txt).

Para evaluar los resultados de la clasificación se utilizará la siguiente matriz de beneficios (o costos):

		Real		
Pronóstico		H	U	L
	H	1	0	-1
	U	0	0,5	0
	L	-1	0	1

La matriz se interpreta de la siguiente manera: si el modelo asigna a un cliente la clase H, y efectivamente pertenece a ella, el resultado obtiene un punto positivo ("+1") para ese vector de datos. Si el cliente fue asignado equivocadamente a la clase L, el resultado obtiene un punto negativo ("-1") para dicho registro. Asignaciones incorrectas de la clase U no serán sancionadas, sin embargo asignaciones correctas de esta clase, serán premiados con 0.5 puntos.

Descripción de los archivos entregados

<i>dmc2004_train.xls</i>	conjunto de entrenamiento, 20.146 clientes, el atributo de la clase no está dado y debe ser generado por usted bajo las condiciones explicadas anteriormente (calcula el Ratio).
<i>dmc2004_desc.xls</i>	detalle de los atributos incluidos en el archivo superior
<i>dmc2004_class.xls</i>	conjunto de 20.146 clientes para los cuales debe ser hecha la clasificación (predicción).

Indicaciones

- La tarea puede ser realizada en grupos de 3 personas.
- Esta tarea será de extensión semestral, la cual tendrá varias entregas que serán avisadas oportunamente y estarán en estrecha relación con los avances del curso.
- Cualquier comentario o consulta, directamente por U - Cursos