UNIDAD TEMÁTICA 7: Ajuste, evaluación y sintonía de modelos

Trabajo de Aplicación 2

ESTUDIO DE CASO – MARKETING BASADO EN AFINIDAD

Un banco crea un nuevo producto financiero, un tipo de cuenta corriente con ciertos costos y tasas de interés, diferentes de otros productos preexistentes. Pasado un tiempo desde el lanzamiento del nuevo producto, una cierta cantidad de clientes han abierto cuentas del nuevo tipo, pero hay muchos otros que aún no lo han hecho.

El departamento de marketing del banco quiere promover las ventas del nuevo producto mediante una campaña de mail directo a los clientes que aún no han optado por el mismo. Sin embargo, con el objetivo de no desperdiciar esfuerzos en clientes que no es probable que compren, desea dirigirse *solamente al* ***20%*** de clientes que tengan la mayor ***afinidad*** por el nuevo producto.

Entonces, ¿cómo podemos determinar si un cliente tiene una gran afinidad por nuestro nuevo producto?

Asumiremos que los clientes que ya han comprado el producto (los compradores) son representativos de aquellos que tienen gran afinidad hacia el mismo. Entonces buscamos clientes que todavía no hayan comprado (los no compradores) pero que sean similares a los compradores en otros aspectos. Nuestra esperanza es que, cuanto más similares sean, mayor será su afinidad.

Nuestro desafío principal es entonces identificar las propiedades de los clientes que nos puedan ayudar a encontrar la similitudes, y que se encuentren disponibles en los datos del banco.

Asumiendo que tenemos buenos datos, podemos utilizar un método de minería de datos estándar para tratar de diferenciar entre compradores y no compradores. Afortunadamente la mayoría de algoritmos pueden generar un *ranking* de clientes, en el cual los que tienen ranking más alto serán clientes predichos como compradores, con mayor nivel de confianza o probabilidad que aquéllos que tengan menor ranking.

Deseamos entonces desarrollar varios modelos de minería, cada uno de ellos capaz de desarrollar un ranking de no compradores en el que los que tengan mayor ranking sean aquéllos para los que el modelo ofrezca más confianza de que deberían, de hecho, ser compradores (si sólo lo supieran!). Veremos también cómo decidir qué modelo es más útil.

Podremos entonces satisfacer al departamento de marketing, proveyéndolo con el 20% superior de los no compradores de nuestro ranking final.

Paso 1 – comprensión del negocio

El banco ofrece cuatro tipos de cuentas corriente, CC01 - CC04, siendo este último el nuevo tipo que se desea promover.

Básicamente, cada tipo de cuenta viene con ciertos costos e intereses mensuales fijos para créditos y débitos, pero algunos clientes pueden tener tasas especiales o estar exentos del pago de los costos mensuales debido a su status VIP u otras particularidades.

Un cliente puede tener cualquier cantidad de cuentas corriente (incluso cero) y cualquier combinación de tipos

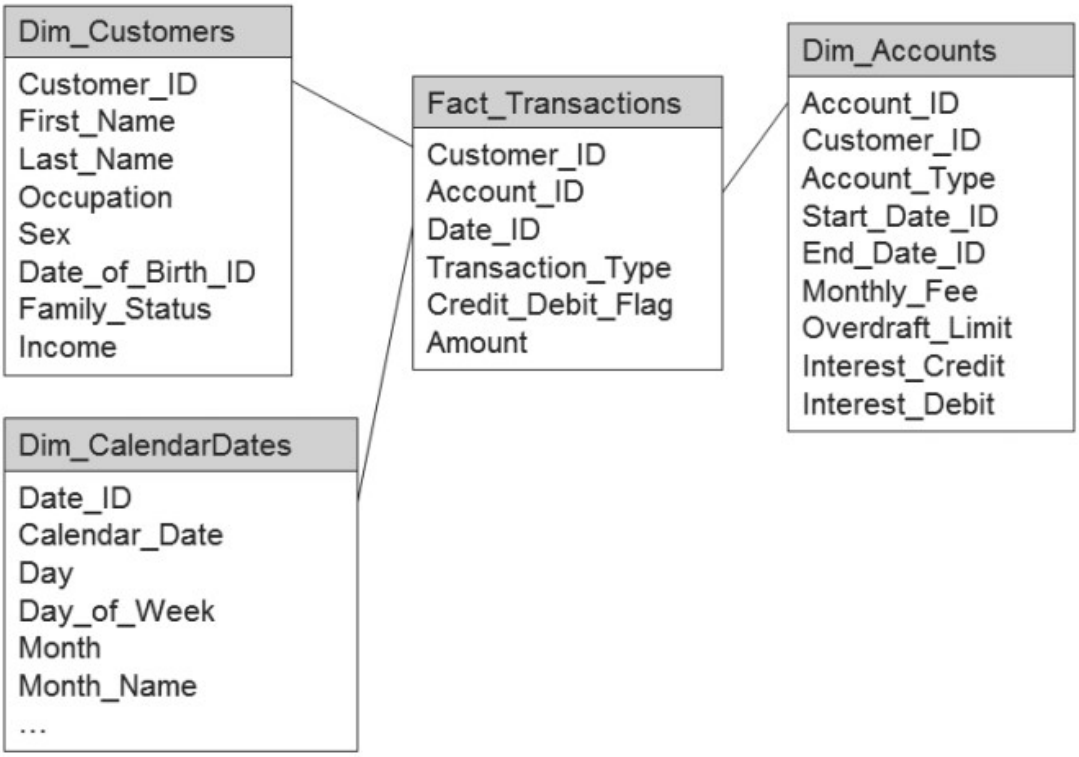
Las cuentas tienen fechas de apertura y cierre. Cuando se cierra una cuenta, su balance es cero y no puede aceptar más transacciones de dinero. Una cuenta abierta cuyo cierre aún no ha pasado, se denomina activa, y un cliente que tiene al menos una cuenta activa es llamado activo.

Cada transacción monetaria, de cada cuenta, es clasificada automáticamente mediante un sistema de análisis de texto interno, basado en un formulario opcional de texto libre que puede ser llenado por el iniciador de la transacción.

Existen varias categorías, como ser "*retiro de caja*", "*sueldo*", "*prima de seguro*", etc., incluyendo una categoría "*desconocido*".

Los datos personales como estado familiar o fecha de nacimiento son conocidos, para la mayoría de los clientes, pero no siempre están actualizados.

Los clientes pueden comprar muchos otros productos del banco, incluyendo cuentas de ahorros, tarjetas de crédito, préstamos o seguros (si bien es muy valiosa, esta información no está disponible en nuestro dataset).


Dim_Customers
Customer_ID
First_Name
Last_Name
Occupation
Sex
Date_of_Birth_ID
Family_Status
Income
Fact_Transactions
Customer_ID
Account_ID
Date_ID
Transaction_Type
Credit_Debit_Flag
Amount
Dim_Accounts
Account_ID
Customer_ID
Account_Type
Start_Date_ID
End_Date_ID
Monthly_Fee
Overdraft_Limit
Interest_Credit
Interest_Debit
Dim_CalendarDates
Date_ID
Calendar_Date
Day
Day_of_Week
Month
Month_Name
…

Paso 2 – comprensión de los datos

En esta etapa corresponde analizar cuidadosamente todos los datos disponibles. Como se mencionó anteriormente, hay varias tablas en el Sistema, en una configuración de Data Warehousing de tipo "Estrella", donde tenemos una "facts table" y varias tablas de dimensiones.

Dim_Customers: una fila para cada cliente, incluye datos personales: nombre y apellido, fecha de Nacimiento (como referencia a la tabla *Dim_CalendarDates*), ocupación), sexo, estado civil, e ingresos. El atributo "income" se pone en cero para los clients jóvenes, pero tiene muchos valores faltantes para los adultos.

• **Dim_Accounts:** una fila para cada cuenta, con una referencia al propietario de la cuenta (customer).

Cada cuenta pertenece a uno de los cuatro tipos **CH01** a **CH04** y tiene fechas de inicio y fin (referencias la tabla *Dim_CalendarDates*).

• **Dim_CalendarDates**: Contiene una fila por cada uno de los 50,000 días 1 de enero de 1900 hasta 22 de noviembre de 2036

• **Fact_Transactions**: Registra todas las transacciones que se han realizado en cada cuenta. Cada transacción incluye una referencia al propietario de la cuenta y la fecha de la transacción . El tipo de la transacción tiene categorías como "salary" , etc (documentar)

El marcador credit/debit (valores *CR* y *DR*) indica si el dinero se ha depositado o extraído de la cuenta. *Amount* es el valor correspondiente, siempre positivo

Se requiere entonces, en esta etapa, analizar y documentar las características de las fuentes de datos, atributos, faltantes, outliers y estadísticas de los atributos disponibles.

Paso 3 – Preparación de los datos

Ver Sección 7.4 del document adjunto.

- Los algoritmos de ML habitualmente utilizan una tabla única integrada y "curada"
- Integración de los datasets de entrada

Generación de un dataset único para aplicar a tareas de ML

- proceso: *01 CreateMiningTable*
  - Ver que se utilizan las 4 tablas como entradas, y se produce una de salida "RawMiningTable"
  - Mostrar el orden de ejecución
  - Tipos de datos…
  - ¿cuáles son los "sub-canales" del proceso existentes, y qué hacen?

Analizar y documentar el resultado emitido por el proceso.

- Sub-stream "customer data"
  - Documentar las operaciones realizadas y su objetivo, así como los operadores utilizados y los parámetros disponibles
  - Sub-stream "account data"
  - ¿Cuál es el objetivo de este proceso?
  - Documentar las operaciones realizadas y su objetivo, así como los operadores utilizados y los parámetros disponibles
- Sub-stream: "transaction data"
  - Ídem
  - Fase final de integración de los sub-streams

Preparación de los datos para minería

- Proceso: "*02 PrepareDataForMining"*
  - Utiliza la salida del proceso anterior
  - Aplica una cadena sencilla de operadores para dejar los datos listos para el modelado
  - Observar el uso del operador *"Set Role"*
  - ¿qué atributo se usa como ID? ¿cómo funciona?
- *"Data cleaning"*
  - valores faltantes: ¿qué se hace? ¿Cuáles serían las alternativas posibles?
  - Atributos inútiles…
- Discretización – analizar y documentar qué se hace y por qué