Universidad de Buenos Aires
Maestría en Minería de Datos y Descubrimiento del Conocimiento
Data Mining en Economía y Finanzas
Edición 2022

Recuperatorio

Fecha límite de entrega domingo 30-abril-2023 a las 23:59:59 ; se aceptan entregas antes de esa fecha debido a que cada alumno tiene un dataset individual distinto al de sus compañeros.
El dataset del recuperatorio consta de clientes de una entidad financiera que poseen la tarjeta de crédito Mastercard.
El objetivo del recuperatorio es construir un modelo predictivo, aplicarlo a datos sin clase, maximizando una métrica de bondad del modelo que se definirá a continuación.

Características del recuperatorio:

- El dataset corresponde a un solo mes.
  - por lo tanto no hay datos históricos
  - al haber un solo mes no se presenta la dificultad del data drifting
  - si se debe hacer Feature Engineering dentro del mismo mes
  - **no** existe el problema de buscar un modelo estable a lo largo de los meses
- por ser tan pocos datos se pueden procesar en la PC local sin recurrir a la nube
- a diferencia del problema planteado en la materia que la clase era ternaria y se podían agrupar las clases, en el recuperatorio la clase es binaria { SI, NO }
- no hay competencia Kaggle, no hay Public Leaderboard ni Private.

Hay dos archivos, en donde <nombre> hace referencia al apellido del alumno

- `<nombre>_generacion.txt.gz`
- `<nombre>_aplicacion.txt.gz`

La primer línea de los archivos tiene el nombre de los campos.
El separador de campos es el <TAB>
El punto decimal es el "."

El archivo `generacion.txt.gz` posee un campo llamado `clase` que admite los siguientes valores

- `SI` clientes que se dan de baja de Mastercard
- `NO` clientes que no se dan de baja

El archivo `generación.txt.gz` es el que se debe utilizar para entrenar el modelo, buscar el modelo que da la mayor métrica de evaluación posible. Obviamente durante el entrenamiento, se debe entrenar en un subconjunto de los registros y testear en otro subconjunto, siempre se debe testear en registros que no fueron vistos por el algoritmo al momento del entrenamiento.

Para estimar que tan bueno es el modelo predictivo se pueden usar las técnicas de :

- Training / Testing
- Montecarlo Cross Validation
- Cross Validation https://en.wikipedia.org/wiki/Cross-validation_(statistics)

El archivo `aplicacion.txt.gz` tiene los mismos campos que `generacion.txt.gz` excepto el campo `clase` .
El archivo `aplicacion.txt.gz` no puede usarse para entrenar el modelo, dado que no tiene clase.
Tampoco se puede utilizar para validar el modelo, ya que no tiene clase.
El archivo `aplicacion.txt.gz` es a donde se debe aplicar el modelo y solo se usan en la etapa final.
Las clientes que están en `generacion.txt.gz` son distintos a las del archivo `aplicacion.txt.gz` .

El campo `numero_de_cliente` es el identificador de los clientes, no se repiten, hay uno por cliente.
Los registros de `aplicacion.txt.gz` son en realidad el 50% tomado al azar para los que intencionalmente se les borro la clase. Son estos registros los que se deben predecir .
Se deben entregar los `numero_de_cliente` del archivo `aplicacion.txt.gz` que a entender del alumno esos registros maximicen esta funcion de ganancia :

Ganancia = 78000 * Aciertos - 2000 * NO_aciertos
Aciertos = SI
No_aciertos = NO

o sea, escrita de otra forma
Ganancia = 78000 * SI - 2000 * NO

Atención : a diferencia de la competencia Kaggle de la materia, aquí la entrega consiste en un archivo de texto cuyo único contenido es una pequeña lista de `numero_de_cliente` , aquellos que a criterio del alumno maximizan la función ganancia. Es la lista de clientes que le pasaríamos al sector de marketing para que haga la campaña de retención de clientes.

La presente función de ganancia da como punto de corte optimo teórico la probabilidad de 0.025, es decir el valor esperado de la ganancia de un registro será mayor a cero si su probabilidad de `clase=SI` es mayor a 0.025
Solamente se debe entregar el campo `numero_de_cliente` .

Por favor, tener en cuenta que solo se deben entregar `numero_de_cliente` del archivo `aplicacion.txt.gz`

Se considerará aprobado el recuperatorio si la ganancia supera a la ganancia de los Ids generados por el script `lineademuerte_recuperatorio.r`
Notar que el script de `lineademuerte` utiliza el algoritmo XGBoost, tambien es posible utilizar su ya conocido LightGBM.

La forma de hacer el trabajo es hacer un modelo predictivo, aplicarlo a los datos sin clase, y entregar los registros con mayor probabilidad de clase=SI que maximicen la función de ganancia.

Podrá enviar hasta un máximo de 5 veces la predicción, la que le será corregida y devuelta. Luego de eso, deberá hacer un nuevo recuperatorio, con un nuevo dataset.