Universidad de Buenos Aires
Maestría en Data Mining
DM EyF
ciclo 2019 "the Red Pill year"

Tarea para el Hogar DIEZ "Herramientas nuevas"

1. Tareas de housekeeping

    - Se han hecho cambios a la carpeta R del dropbox, se han agregado scripts y se han modificado scripts existentes, bajarla nuevamente a la PC local y luego subir al bucket de Google Cloud .
    - Mi recomendación, es que cada dia que vayan a correr scripts, se bajen nuevamente los que estan en la nube, ya que estamos en un periodo de gran efervescencia.

2. Observación de un ayudante en otra universidad

" b. Respecto a la creación de nuevas variables que incluyeran medias, ratios móviles y ponderaciones mensuales (mayor peso al último mes) hubo un desconocimiento generalizado de parte de los alumnos en cuanto a la dificultad técnica de lograr tal acometido.

En otras palabras, aún no dieron el salto técnico para desarrollar nuevas funciones que realicen loops sobre las columnas (aplicado al dataset grande, premium).
Aún menos tocar siquiera una coma del código en C de fe_todoenuno.

En resumen, ví entusiasmo en cuanto a la creación teórica de qué variables les gustaría crear pero cierto desconocimiento en el cómo hacerlo."

3. Leer https://www.crondose.com/2016/09/developer-learning-curve/
   No se sienta frustrado si lleva menos de 300 horas programando en R.
   Notará cambios enormes al pasar las 1.000 horas en un próximo futuro del año 2020.

4. Nueva version de script `fe_todoenuno.r`

Se ha modificado el script `R/FeatureEngineering/fe_todoenuno.r` para agregar el cálculo de le media movil de los ultimos 6 meses.
Este nuevo campo aparece con la extencion `__avg`

Esto NO modifica la forma en que se construyó la linea de muerte, que solo usa los campos `__min, __max y __tend`

5. Nuevo script `lightgbm_directo_wfv_auto.r`

Anteriormente trabajabamos con dos probabilidades fijas.
La primera es de 0.025 que utilizamos cuando entrenamos sobre el 100% del dataset
La segunda es de 0.2040816 = 10*500 / (19500 + 10*500) que es cuando usamos un undersampling del 10% de los negativos={BAJA+1, CONTINUA} , el 10 surge de hacer 1/0.1 (0.1 proviene del 10%)

Ahora, se ESTIMA la probabilidad de corte en el dataset de undersampling, y luego en forma teórica se calcula cuanto tiene que ser esa probabilidad en el dataset normal del 100% de los datos

Tan solo explicar lo que esta haciendo la nueva funcion `fganancia_logistic_lightgbm` , con el oscuro uso de los weights como artificio para hacer un validation dentro de los datos de testing, donde en ese validation se estima la mejor probabilidad de corte, llevaria al menos 1 hora de pizarron.
En caso que un grupo *significativo* de alumnos considere la próxima clase que no pueden continuar adelante con sus vidas sin entender inmediatamente esa parte del código, será explicado en clase.
Caso contrario, utilizaremos el tiempo para ver temas de como seguir sumando ganancia.

Con este script se obtienen ganancias interesantes.

6. Nuevo script `lightgbm_directo_wfv_auto.r`

Este script trabaja con este concepto en entrenamiento
positivos = { BAJA+2, BAJA+1 }
negativos = { CONTINUA }

Atencion, se entrena suponiendo positivos = { BAJA+2, BAJA+1 } PERO siempre al momento de evaluar la ganancia se recurre a la formula original donde los BAJA+2 valen $ 19.500 y los BAJA+1 y CONTINUA valen $ -500

En este caso optimizar la probabilidad de corte es algo indispensable, ya que por lo general para de los 0.025 a 0.036 / 0.040 en el dataset sin undersampling.

Tiene toda la complejidad del script anterior

Muy estimado Federico, dado que he corrido el script y las ganancias que obtengo NO son buenas, he tomado la decisión pedagógica este año de no agregar código para trabajar con BAJA+3 y/o BAJA+4
Quizas, este año hay que buscar el oro en otros lados de la montaña.