ITBA
Especialización en Ciencia de Datos
Data Mining 2022A

Tarea para el Hogar TRES

El objetivo fundamental de esta tarea para el hogar es que usted se familiarice con la corrida de scripts en el entorno Google Cloud, entienda la dinámica de las corridas, donde están los scripts, datasets, donde quedan los resultados, cómo es el funcionamiento de git, GitHub, se enfrente a los problemas más comunes, se le apaguen máquinas virtuales y deba relanzar procesos.
Si no posee conocimientos de informática al comienzo le parecerá todo muy caótico, desconexo, innecesariamente complejo. Luego de diez corridas, ya sabrá como es el mecanismo. Luego de dos semanas comprenderá la razon de ser de los mecanismos.
En este arduo camino, continuará escalando la colina, lo que le dará fuerzas para no abandonar.

Deberá estar muy atento realizar experimentos que terminen en un tiempo razonable.
Con el feature engineering histórico deberá ser muy cuidadoso con la cantidad de campos finales que posea su dataset, ya que datasets muy grandes demandarán decenas de horas en la optimización bayesiana.
Adicionalmente, elegir la cantidad de meses en las que va a entrenar afectará dramáticamente los tiempos de corrida de la optimización bayesiana.

Esta es la penúltima tarea para el hogar, aún no se está corriendo con lo más avanzado de la asignatura. En particular aún falta la metodología MLOps, con lo cual notará cierta desprolijidad en la forma de correr los scripts, lo que le demandará una ferrea disciplina de registración de los experimentos.
Lamentablemente se estará utilizando 5-fold cross validation en la optimización bayesiana para la optimización de los hiperparámetros del lightgbm, lo que hará que las corridas sean muy largas; ármese de paciencia.
Se entrenará en todo un conjunto de meses y no se utilizará la opción de hacer subsampling de la clase CONTINUA para acelerar la optimización bayesiana.
Tampoco se estarán utilizando los métodos se *semillerio* ni de *hibridación de semillerios* .

Sea paciente, los momentos gloriosos ya llegarán.

1. Corrida Motivacional

Correr `src/TareaHogarTRES/z652_lightgbm_motivacional.r`

- El script debe correr en Google Cloud, ya que utiliza el dataset con todos los meses
- Lea al comienzo del script los requerimentos de Virtual Marchine
- La salida del script queda en el Google Storage Bucket, recuerde que accede por https://console.cloud.google.com/storage/browse, en la carpeta ./exp/KA6520
- La corrida genera varios archivos .csv , bájelos a todos a su computadora local, y luego haga el submit de cada uno a Kaggle
- También genera un archivo con la importancia de variables, analícelo.
- Se sugiere que la primera vez lo corra tal cual está, y luego realice cinco corridas más, cada una de ellas cambiando la semilla, en la línea 88
- si se siente en un día intrépido, y está decidido a tomar riesgos, pruebe modificar los hiperparámetros de la llamada a LightGBM
- lea el script en gran detalle para entender que es lo que está haciendo

2. Perspectiva de la "escalada completa de la colina"

Esta es la última vez que usted entrenará utilizando un solo mes, 202011 .
Las celdas en color naranja ya las calculó en la Tarea para el Hogar DOS, copie esos números aquí nuevamente.
Ahora deberá completar nuevas doce celdas nuevas de la siguiente tabla realizando las corridas necesarias , algunas de estas corridas serán en la nube y otras en la PC local, deberá leer las primeras líneas del script para saber donde correrlo.
Todos los alumnos obtendrán tablas distintas ya que utilizarán sus propias semillas.

Tabla Escalando la Colina		
Método	**Ganancia cross validation/Montecarlo**	**Public Leaderboard**
rpart default		
rpart con hiperparámetros óptimos		
ranger default		
ranger con hiperparámetros óptimos		
lightgbm default		
lightgbm		

con hiperparámetros óptimos		
xgboost original default		
xgboost original con hiperparámetros óptimos		
xgboost histograma default		
xgboost histograma default con hiperparámetros óptimos		

El "XGBoost original" es tal cual el del paper original de Tianqui Chen del año 2016
En el año 2017 Microsoft realiza una "copia con mejoras" que denomina LightGBM, siendo la principal mejora no trabajar con los atributos continuos sino discretizarlos antes, binning, método que llamaron de histogramas.
Tan exitosa fue la mejora de LightGBM que XGBoost adopta en 2018 los histogramas.
XGBoost Histograma funciona varias veces más rápido que XGBoost Original

- **lightgbm con hiperparámetros óptimos**, realice la corrida del script `labo/src/lightgbm/z533_lightgbm_BO.r` (demandará varias horas) la salida queda en `labo/exp/HT5330/HT533.txt`
  - del archivo, la mayor ganancia es la que debe poner en la columna "Ganancia cross validation/Montecarlo" de la tabla.
  - Con los hiperparámetros de la mejor ganancia vaya al script `labo/src/lightgbm/z512_lightgbm.r` , reemplace por esos hiperparámetros entre a partir de la línea 31 (quizás deba agregar alguna línea), córralo, lea en el script donde queda la salida, suba la salida a Kaggle, y de allí obtendrá el valor que debe poner en la columna "Public Leaderboard" de la tabla.

- **lightgbm default**, corra el script `labo/src/lightgbm/z531_lightgbm_default.r` con su propia semilla (correrá en pocos minutos) y dejará la salida en `labo/exp/HT5310`
  - la salida se mostrara por pantalla y la deberá cargar en la columna "Ganancia cross validation/Montecarlo" de la tabla.
  - Con los unicos hiperparámetros del archivo (que son los default de lightgbm) vaya al script `labo/src/lightgbm/z512_lightgbm.r` , reemplace por esos hiperparámetros alrededor de la línea 31, córralo, suba la salida a Kaggle, y de allí obtendrá el valor que debe poner en la columna "Public Leaderboard" de la tabla.

- **xgboost orignal con hiperparámetros óptimos**, realice la corrida del script `labo/src/xgboost/z563_xgboost_original_BO.r` (demandará una gran cantidad de horas) la salida queda en `labo/exp/HT5630/HT563.txt`

    - del archivo, la mejor ganancia es la que debe poner en la columna "Ganancia cross validation/Montecarlo" de la tabla.
    - Con los hiperparámetros de la mejor ganancia vaya al script `labo/src/xgboost/z561_xgboost_original.r` , reemplace por esos hiperparámetros alrededor de la línea 31 sin olvidar el nrounds, córralo, lea en el script donde queda la salida, suba la salida a Kaggle, y de allí obtendrá el valor que debe poner en la columna "Public Leaderboard" de la tabla.

- **xgboost original default**, corra el script `labo/src/xgboost/z562_xgboost_original_default.r` con su propia semilla (correrá en pocos minutos) y dejará la salida en `labo/exp/HT5620`
    - del archivo, a la ganancia la deberá cargar en la columna "Ganancia cross validation/Montecarlo" de la tabla.
    - Con los unicos hiperparámetros del archivo (que son los default de xgboost) vaya al script `labo/src/xgboost/z561_xgboost_original.r` , reemplace por esos hiperparámetros alrededor de la línea 31 sin olvidar nrounds, córralo, suba la salida a Kaggle, y de allí obtendrá el valor que debe poner en la columna "Public Leaderboard" de la tabla.

- **xgboost histograma con hiperparámetros óptimos**, realice la corrida del script `labo/src/xgboost/z573_xgboost_histograma_BO.r` (demandará una gran cantidad de horas) la salida queda en `labo/exp/HT5730/HT573.txt`
    - del archivo, la mejor ganancia es la que debe poner en la columna "Ganancia cross validation/Montecarlo" de la tabla.
    - Con los hiperparámetros de la mejor ganancia vaya al script `labo/src/xgboost/z571_xgboost_histograma.r` , reemplace por esos hiperparámetros alrededor de la línea 31 sin olvidar el nrounds, córralo, lea en el script donde queda la salida, suba la salida a Kaggle, y de allí obtendrá el valor que debe poner en la columna "Public Leaderboard" de la tabla.
    -
- **xgboost histograma default**, corra el script `labo/src/xgboost/z572_xgboost_histograma_default.r` con su propia semilla (correrá en pocos minutos) y dejará la salida en `labo/exp/HT5720`
    - del archivo, a la ganancia la deberá cargar en la columna "Ganancia cross validation/Montecarlo" de la tabla.
    - Con los unicos hiperparámetros del archivo (que son los default de xgboost) vaya al script `labo/src/zxgboost/z571_xgboost_histograma.r` , reemplace por esos hiperparámetros alrededor de la línea 31 sin olvidar nrounds, córralo, suba la salida a Kaggle, y de allí obtendrá el valor que debe poner en la columna "Public Leaderboard" de la tabla.

Una vez completada la nueva y extendida tabla "Escalando la colina" compártala en Zulip.

3. Agregando hiperparámetros a LightGBM

Esta tarea implica programación en R, modificar un script ; preste mucha atención.

Leer
https://lightgbm.readthedocs.io/en/latest/Parameters-Tuning.html
http://devdoc.net/bigdata/LightGBM-doc-2.2.2/Parameters-Tuning.html
https://neptune.ai/blog/lightgbm-parameters-guide

Finalmente, hacer una copia del script `labo/src/lightgbm/z543_lightgbm_BO_nube.r` y agregar a la optimización bayesiana los hiperparámetros a optimizar que desee, hasta ahora las corridas que ha efectuado utiizan los valores por default de los hiperparámetros que no están siendo incluidos en la optimización bayesiana.

Como ejemplo tiene disponible el script `labo/src/lightgbm/z537_lightgbm_BO_hyper.r` en donde se han agregado los hiperparámetros `lambda_l1` y `lambda_l2`, compare en detalle los scripts 543 y 537 así se las ingenia para modificar el script y agregar los hiperparámetros que usted considera van a permitirle aumentar la ganancia.
Notar que

1. se pasó de 100 iteraciones a 150 iteraciones (línea 20) ya que el espacio de búsqueda es más grande
2. en la línea 28 se agregan los hiperparámetros `lambda_l1` y `lambda_l2`
3. alrededor de la liínea 95 se comentan con `# lambada_l1` y `lambda_l2` ya que ahora son hiperparámetros
4. obviamente, alrededor de la línea 147 se cambia la carpeta donde va el resultado

4. Generacion de Modelos

Este ejercicio es el que más creatividad le demandará, y podrá realmente escalar la colina si realiza varios experimentos, que correrá en paralelo, con varias virtual machines encendidas al mismo tiempo en Google Cloud, y va aprendiendo cual es la combinación que genera mejores variables.

Utilizará básicamente tres scripts :

- `src/FeatureEngineering/z601_FE_historico.r`
- `src/lightgbm/z602_lightgbm_BO_xval.r`
- `src/lightgbm/z603_lightgbm_final.r`

`src/FeatureEngineering/z601_FE_historico.r`

Es el script más rico, que le permitirá desplegar todo su creatividad, espíritu experimental, y para muchos será una batalla contra el empiricismo, abra su mente a nuevas ideas, acepte que lo que hoy usted llama "crear variables que tienen sentido" es en realidad sentido en relación a su concepción del mundo o a teorias desarrolladas por otros, que pueden estar mal, o ser imperfectas. Abrace el descubrimiento, abrace la maravilla de la invención, de la sorpresa; para seguir cómodo se quedaba siendo ingeneiro o economista bajo la sombra gris de los protocolos.
Operación, en cada corrida deberá determinar el nombre del archivo de salida (línea 22) , en la parte donde se inicia el programa debe elegir que funciones quiere efectivamente llamar y cuales comentar para que no se ejecuten.
Siéntase libre de agregar todas las variables nuevas que desee en la función AgregarVariables()
Sea agresivo y agrege más tendencias, ya no con una ventana de 6, sino con una mas larga, de 12 por ejemplo, o de 3.
Agregue lagsde orden 2 o superiores.
Descomente las lineas del Rankeador
Recuerde que simpre podrá llamar a CanaritosImportancia() que le reducirá la cantidad de variables, para que queda un dataset manejable.
Hint avanzado, ¿quiere ser realmente agresivo? cambie agregue en cada paso todas las columnas a `cols_lagueables`

`src/lightgbm/z602_lightgbm_BO_xval.r`

Decida con sagacidad en que periodo entrenar.
No tiene intuición en este nuevo dominio bancario que le es ajeno?, experimente con distintos períodos, descubra como funciona este nuevo mundo. Entienda que es lo que sucedió en los meses de pandemia en ese mundo.
Si realizó la tarea de "Agregando hiperparámetros a LightGBM" y obtuvo alentadores resultados, incorpore esos hiperparámetros a este script.
Operación, debe cargar al inicio el nombre del experimento, asigne correctamente el nombre del dataset que se utilizará, que debe ser el generado en el Feature Engineering.

`src/lightgbm/z603_lightgbm_final.r`
Diga adiós a la probabilidad de corte, pruebe en el Public Leaderboard la cantidad de envios, 1's en el archivo de salida, que le conviene enviar.
LightGBM devuelve probabilidades descalibradas muchas veces, no se guie por ellas, pase directamente a la cantidad de estímulos. La probabiidad de 1/60 sigue siendo válida, es Lightgbm quien las está devolviendo descalibradas, el orden está bien, la probabilidad no.
Operación : debe cargar al inicio el nombre del experimento, nombre del dataset de entrada, los meses en los cuales se va a entrenar el modelo final, y debe copiar a mano los mejores hiperparámetros encontrado en la Bayesian Optimization

5. Lecturas

- Para quien aún no termina de entener que es esto de los modelos predictivos, qué es la clase, un ejemplo deliciosamente escrito, no técnico, de como se construye y aplica un modelo en una empresa (una novela ligera, 40 minutos de lectura)
  - https://storage.googleapis.com/dmeyf/annotation/general/Target.html

- Lecturas cortas, especialmente dedicadas a Bianca Balzarini (pero abiertas a todos)
  - https://storage.googleapis.com/dmeyf/annotation/general/why_business_fail_at_machine_learning.html
  - https://storage.googleapis.com/dmeyf/annotation/general/why_so_many_data_scientists_are_leaving_their_jobs.html

- Lectura sobre donde poner la energía
  - https://storage.googleapis.com/dmeyf/annotation/general/MostMisunderstoodHero.html

- Sobre Valores Shapley
  - https://www.youtube.com/watch?v=ngOBhhINWb8
  - https://www.youtube.com/watch?v=B-c8tIgchu0

- Inferencia de Causa Efecto
  - https://www.youtube.com/watch?v=Od6oAz1Op2k
  - https://www.youtube.com/watch?v=dFp2Ou52-po

6. Hoja de ruta para las clases del 03 y 04 de junio

Hoja de ruta de las clases 7 y 8

Clase 7, viernes 03 de junio

17:30 a 18:45 Sesión 1 laboratorio

- Discusión de los resultados de la Tarea para el Hogar TRES
- Metodología MLOPs de experimentación en Google Cloud, presentación del entorno MLFlow y experimentos en Google Cloud

18:45 a 19:00 Break

19:00 a 20:20 Sesión 2 Exposición

- Operación de MLOPs y MLFlow
- Análisis Sesgo - Varianza
- Método de Ensemble : Semillerios

20:20 a 20:35 Break

20:35 a 22:00 Sesión 3 laboratorio

- Método de Ensemble : Hibridación de Semillerios

Clase 8, sábado 04 de junio

09:00 a 10:15 Sesión 3 laboratorio

- Análisis de la importancia de variables utilizada hasta el momento
- Importancia de variables con la metodología de Valores Shapley

10:15 a 10:30 Break

10:30 a 11:50 Sesión 4 laboratorio

- Métrica de distancia utilizando Random Forest
- Clustering Jerárquico
- Tips para video a Miranda Wintour

11:50 a 12:10 Break

12:10 a 13:30 Sesión 4 laboratorio

- Sacando ventaja del Public Leaderboard, como elegir la cantidad de envios que va a optimizar el Private