Universidad Austral
Maestría en Ciencia de Datos
Laboratorio de Implementación I , 2024 virtual v1

Fecha de liberación oficial: martes 09 de abril 23:01

Tarea para el Hogar CUATRO

Esta Tarea para el Hogar es un punto de quiebre, es la última en donde entrenaremos en un solo mes 202107, a partir de la *Clase 05* empezaremos a utilizar la historia.

Todo lo que hacemos es para mejorar la ganancia de la predicción en los datos nuevos; los Experimentos Colaborativos tienen como objetivo mejorar la ganancia, decididamente no son experimentos académicos y mucho menos su espíritu será comparar dos alternativas que funcionan ambas mal.

Es fundamental que usted lea en el recientemente actualizado Libro de la Asignatura el capítulo 6 sobre Experimentos Colaborativos, luego vaya al documento compartido de Google Slides de Experimentos Colaborativos, lea muy detenidamente en que consisten, forme grupo, y elija un experimento que sea de su interés y esté a su alcance.

Si su computer literacy actualmente es bajo, desde la cátedra le sugerimos fuertemente que para Experimentos Colaborativos busque formar grupo con alguien con mayores habilidades. Este es el consejo más práctico que recibirá de la cátedra en las ocho clases, ahórrese innecesarios padecimientos.

Sección Pasado (ya lo debería haber hecho)

1. Videos Pasado (ya los debería haber visto)
   ver los videos que aún no ha visto en el campus virtual, Clase 04
   1. Workflow de trabajo
   2. Catastrophe Analysis
   3. Data Drifting
   4. Feature Engineering Intra Mes
   5. Feature Engineering Histórico

   tiempo humano estimado: **45** minutos (a 1.5x)

2. Análisis Variables Rotas

Analizará un problema que el sector de DataWarehouseing ha tenido en la generación de los datos, tal cual se mostró en el video Catastrophe Analysis

Desde la máquina virtual `desktop` que está en Sao Paulo corra el script `./src/CatastropheAnalysis/z505_graficar_zero_rate.r`

El proceso demorará alrededor de 20 minutos

La salida del script queda en su bucket, en la carpeta `~/buckets/b1/exp/CA5050`
Analice los archivos, en particular `zeroes_ratio.pdf`

tiempo computacional: **20 minutos**
tiempo humano estimado : **15 minutos**
dificultad : **baja**
creatividad requerida : 1**0%**

Sección Deseable

3. Corrida inicial `ranger`

Este script correrá con sus semillas (ya almacendas en el bucket) y le dará resultados distintos al de sus compañeras/os
La libreria `ranger` es una implementación del algoritmo Random Forest. Por favor no confunda Random Forests el famoso algoritmo creado por Leo Brieman en 2001 con el casero Arboles Azarosos
Arboles Azarosos solo randomiza las columnas que se utilizan para la construccion de cada árbol
Random Forest randomiza los registros del dataset con bootstrapping https://en.wikipedia.org/wiki/Bootstrapping_(statistics) y en además cada split de cada arbol utiliza, en forma random, apenas un subconjunto de los atributos.
Corra el script `src/ranger/z431_ranger.r`
y suba a Kaggle la salida generada `~/buckets/b1/exp/KA4310/KA4310_001.csv`

Por favor no se angustie ante magras ganancias en Kaggle de este script inicial, en el próximo ejercicio hará una Optimización Bayesiana para encontrar los hiperparámetros óptimos de `ranger` para nuestro dataset.

tiempo de corrida **< 5 minutos**
tiempo humano : **3 minutos**
dificultad : **muy baja**

4. Corrida Bayesian Optimization `ranger`

   Este script correrá con sus semillas (ya almacendas en el bucket) y le dará resultados distintos al de sus compañeras/os

   Corra el script `src/ranger/z433_ranger_BO.r`

   tiempo de corrida **+8 horas**
   tiempo humano : **30 minutos**
   dificultad : **baja**

5. Corrida `ranger` con hiperparámetros óptimos

   Este script correrá con sus semillas (ya almacendas en el bucket) y le dará resultados distintos al de sus compañeras/os

   Levante a una planilla la salida del ejercicio anterior, que ha quedado en `~/buckets/b1/exp/HT4330/HT4330.txt`, ordene en forma descendente por la ganancia, y copie los mejores hiperparámetros en en script `z431_ranger.r` para finalmente generar la salida para Kaggle.

   Cargue los resultados del primero del ranking en la Planilla Colaborativoa solapa `ranger`, tanto para la ganancia que aparece en el archivo de salida como para el valor del Public Leaderboard

   ¿Cómo comparan el mejor resultado del renombrado algoritmo Random Forest con la mejor corrida colaborativa de Arboles Azarosos, ambas en el Public Leaderboard?

   Según los resultados obtenidos por ranger, y los que obtenga más adelante en la Tarea para el Hogar con la corrida de LightGBM, usted decidirá si vale la pena entender el script en detalle.

   tiempo de corrida **< 5 minutos**
   tiempo humano : **5 minutos**
   dificultad : **baja**

6. Corrida Inicial LightGBM

Este script correrá con sus semillas (ya almacendas en el bucket) y le dará resultados distintos al de sus compañeras/os
Corra el script `src/lightgbm/z454_lightgbm_final.r`
y suba a Kaggle las varioas salidas generadas en la carpeta `~/buckets/b1/exp/KA4540/`

tiempo de corrida **< 5 minutos**
tiempo humano : **5 minutos**
dificultad : **muy baja**

7. Corrida Bayesian Optimization LightGBM

Este script correrá con sus semillas (ya almacendas en el bucket) y le dará resultados distintos al de sus compañeras/os
Corra el script `src/lightgbm/z451_lightgbm_binaria_BO.r`

tiempo de corrida **+8 horas**
tiempo humano : **3 minutos**
dificultad : **muy baja**

8. Corrida LightGBM con hiperparámetros óptimos

Este script correrá con sus semillas y le dará resultados distintos al de sus compañeras/os
Levante a una planilla la salida del ejercicio anterior, que ha quedado en `~/buckets/b1/exp/HT4510/HT4510.txt`, ordene en forma descendente por la ganancia, y anote los hiperparámetros óptimos:
Modifique el script `src/lightgbm/z451_lightgbm_final.r` para que que grabe el resultado en una nueva carpeta de experimento, y cambie los hiperparámetros a los encontrados en el punto anterior. Finalmente suba las salidas generadas a Kaggle.
Cargue los resultados del primero del ranking en la Planilla Colaborativoa solapa `LightGBM`, tanto para la ganancia que aparece en el archivo de salida como para el valor del Public Leaderboard

tiempo de corrida **< 5 minutos**
tiempo humano : **5 minutos**
dificultad : **baja**

En caso que LightGBM sea su mejor modelo predictivo, lea en detalle ambos scripts de LightGBM, entendiendo linea a linea el código
tiempo humano : **60 minutos**
dificultad : **alta**

9. Lecturas sobre overfitting & Private Leaderboard

- Overfitting the Leaderboard in Ernst & Young Data Science Competition 2019
  And subsequently losing 8000 USD + a ticket to New York.

  https://medium.com/hmif-itb/overfitting-the-leaderboard-da25172ac62e

- The dangers of overfitting: a Kaggle postmortem https://gregpark.io/blog/Kaggle-Psychopathy-Postmortem

tiempo humano : **15 minutos**

10. Lectura de Reglas de Experimentos Colaborativos

Lea de El Libro de la Asignatura el capítulo "6 Experimentos Colaborativos"

tiempo humano estimado: **10 minutos**

11. Lectura de Google Slides de Experimentos Colaborativos

Se ha actualizado recientemente *El Libro de la Asignatura*, en el capítulo **1.2 Links Fundamentales** se ha agregado el link de `Google Slides Experimentos Colaborativos`, ingrese al mismo y solicite permisos, a las pocas horas será autorizado.
Una vez que reciba el email notificándole que ya tiene acceso al Google Slides de Experimentos Colaborativos, léala, reúnase con los compañeros con los que suele hacer grupo, y decida con cual experimento va a participar. Escriba su nombre en la portada del experimento.
Negocie en Zulip con el resto del curso en caso de haber colisiones.

tiempo humano estimado **30 minutos**

A continuación encontrará, en orden alfabético, la lista de alumnas/os que por sus sesgos cognitivos previos, a fines pedagógicos la cátedra ya les ha pre asignado un experimento. En caso de querer optar por otro experimento, estos estudiantes deberán ponerse en contacto con la cátedra.

Estudiante	Experimento pre-asignado
Salinas, Cristian	Problema #03 Feature Engineering ¿Reducir la dimensionalidad del dataset mejora el modelo predictivo? o sorprendentemente lo empeora?

https://www.youtube.com/watch?v=68t74jofBD0

12. Videos Prioritarios clase 05

ver los videos que aún no ha visto de la clase 04 y 05

1. Training Srategy
2. Hyperparameter Tuning
3. Etapas Finales

tiempo humano estimado si debe verlos todos: **35 minutos** (a 1.5x)

Sección Complementaria

13. Mejoras al script de Optimización Bayesiana de LightGBM

Este el el ejercicio más interesante de toda esta Tarea para el Hogar
Lea https://neptune.ai/blog/lightgbm-parameters-guide o busque algún artículo que le muestre cuales son los hiperparámetros m´s importantes de LightGBM
Haga una copia del script `src/lightgbm/z451_lightgbm_binaria_BO.r` para que tengan en cuenta en la Optimización Bayesiana esos hiperparámetros que encontró en la literatura, luego corra el script y siga el camino habitual hasta ver la ganancia en Kaggle

tiempo de corrida **+8 horas**
tiempo humano : **60 minutos**
dificultad : **muy, pero muy alta . No lo intente en ayunas**