ITBA
Maestría en Ciencia de Datos
Data Mining, 2024 A

Fecha de liberación oficial: viernes 24 de mayo 08:00

Tarea para el Hogar TRES

Esta es la última Tarea para el Hogar, hará la primera corrida de Experimentos Colaborativos con el Wokflow Baseline. Este corrida demandará diez horasde procesamiento, y le permitirá concentrarse en los resultados de SU experimetno colaborativo.

En Experimentos Colaborativos alumnos organizados en grupos optimizarán distintas etapas del workflow y colectivamente se encontrarán las mejores prácticas, que se aplicarán directamente a la entrega final de la Competencia Kaggle intentando lograr la mayor ganancia posible.

No se angustie si por primera vez no es sorprendido con ganancias superadoras, la sábana es corta, el Workflow Baseline fue pensado para que usted pueda superarlo.

Le llevará las clases 07 y 08 terminar de entender conceptualmente la funcionalidad de cada etapa del workflow, sus parámetros y la forma en que cada etapa afecta a las que siguen. La determinación de los parámetros óptimos de cada etapa será empírica ya que depende de las caracteristicas de estos datos y de la correlación entre las variables independientes y la clase.

Es la idea que los alumnos que superaron la etapa de la angustia prueben configuraciones distintas de los parámetros del Workflow Baseline, solo podrá destacarse del resto si deja de ser obediente.

Sección Deseable

1. Videos Sobre el Workflow

Aqui hay videos anteriores que ya debería haber visto y videos nuevos, páselos rápido, evite morir de aburrimiento.

- Etapas Generales
 - o Pres Etapas Generales
 - Video Etapas Generales
- Catastrophe Analysis
 - Pres Catastrophe Analysis
 - Video Catastrophe Analysis
- Data Drifting

Augusto, aquí está explicado "ranking con cero fijo"

- Pres Data Drifting
- Video Data Drifting
- Feature Engineering IntraMes
 - Pres Feature Engineering IntraMes
 - <u>Video FeatureEngineering IntraMes</u>
- Feature Engineering Histórico
 - o Pres Feature Engineering Histórico
 - Pres Feature Engineering Histórico
- Training Strategy
 - Pres Training Strategy
 - Video Training Strategy
- · Hyperparameter Tuning
 - Pres Hyperparameter Tuning
 - Video Hyperparameter Tuning
- Etapas Finales
 - o Pres Etapas Finales
 - Video Etapas Finales

2. Lectura instrucciones Experimentos Colaborativos Lea el capítulo 6 "Experimentos Colaborativos" de El Libro de la Asignatura.

3. Lectura Google Slides de Experimentos Colaborativos

Lea el documento de Google Slides compartido en la cartelera de Zulip, deberá solicitar acceso . Allí es donde usted documentará su experimento.

4. Registración en Experimentos Colaborativos

Anote su nombre en la carátura del Experimento Colaborativo que más le interese. En su vida profesional de Data Scientist va a tener que trabajar en todas las etapas todos los meses, así que lo que no elija hoy lo terminará haciendo.

5. Corrida del Workflow Baseline

De lo que entendió en clase y en los videos sobre el workflow, identifique cual etapa del workflow se corresponde a su Experimento Colaborativo

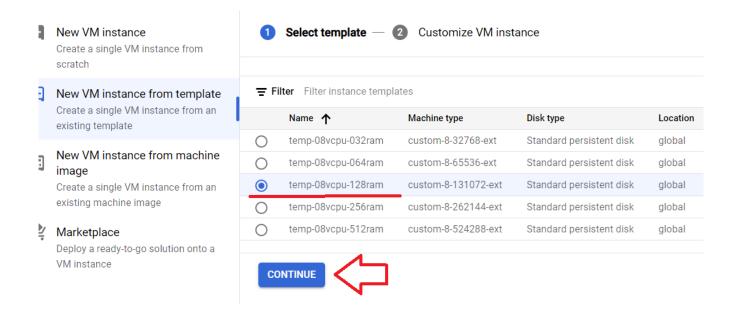
Actualice su repositorio GitHub

Junto a su compañera/o de equipo lea en gran detalle el script ./src/workflow01/z507_workflow_base.r

Haga una copia del mismo, quitándole la letra "z" inicial

Cada inegrante del grupo debe actualizar su repositorio, subiendo su nuevo script modificado a GitHub

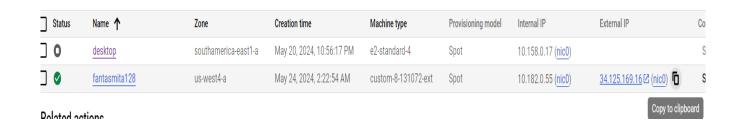
Vayan a la consola de Google Cloud https://console.cloud.google.com/compute/instancesAdd? &creationFlow=fromTemplate y marque la opcion temp-08vcpu-128ram para crear una virtual machine de 128 GB de memoria RAM y 8 vCPU



Presione el boton Continue

Dentro de la pantalla de configuración, llame a esa virtual machina fantasmita128, y elija como región Las Vegas y como zona us-west4-a

vaya al fondo de la pantalla y presione el botón azul de CREATE Espere a que su virtual machine fantasmita128 se encienda y copie la External IP al portapapeles



Desde un browser en su PC ingresa a http:://(la ip externa que copio)

```
por favor tenga en cuenta que es http SIN ninguna "s" al final En el ejemplo de la imagen, se debio escribir http://34.125.169.16/
```

Le aparecerá la pantalla de credenciales para ingresar a RStudio, cargue su usuario y su clave.

Ya dentro de RStudio , busque SU script al que le quitpo la "z" inicial y si se animó, modificó ~/dm2024a/src/workflow01/z509 workflow base.r

Ponga a correr su script, directamente con el botón Soruce() de Rstudio

La corrida demorará 10 horas

Este profesor le desea que Google Cloud le mate su spot virtual machine fantasmita128, así forja su espíritu y desarrola la resiliencia ya que la necesitará en abundantes cantidades para las corridas de Experimentos Colaborativos. Si esto sucediera, no entre en pánico, simplemente ponga a correr nuevamente el script.

En caso que se le desconecte RStudio y que también falle el intento de reconexción, NO se desespere, ingresa a una terminal Ubuntu de esa máquina y corra el comando <a href="https://

Al terminar la virtual machine fantasmita128 se auto.suicidará y desaparecerá. Esa será la forma de saber que terminó.

Los resultados de los experimentos de la corrida quedan en el bucket, dentro de la carpeta expw, en las siguientes carpetas

- DT-0001
- CA-0001
- FEintra-0001
- DR-0001
- FEhist-0001
- FErf-0001
- TS-0001
- HT-0001
- ZZ-0001

Por favor, interprete también lo que queda en el bucket en la carpeta flow

Analice lo que quedó en cada una de las carpetas, olfatee cada archivo.

Analice con su grupo lo que quedó en la carpeta correspondiente a la etapa del Experimento Colaborativo que eligió

Finalmente, el paso más importante de todos, lea en grupo detalladamente el script correspondiente al experimento colaborativo elegido

Anexo

El workflow se comunica con usted con mensajes directos de Zulip desde el inframundo y con archivos de control que deja en la carpeta del experimento **expw**

- parametros.yml son los parámetros que recibió el script, los que usted configuró en el workflow.
- z-Rbegin.txt indica que el script R de su experimento comenzó.
- output.yml quedan relevantes resultados de la corrida del script
- z-Rend.txt indica que la corrida del script de R fue completamente exitosa.
- z-SHlog.txt queda la fecha y hora en que se inició el script, y en el caso que terminó exitosamente está la fecha de finalización
- z-SHoutfile.txt posee todo lo que fue saliendo por pantalla mientras corria su script, le servirá en caso que abortara para investigar el origen de error.
- z-Rabort.txt indica que el R ha abortado. El contenido del archivo z-SHoutfile.txt le podrá dar una pista del error.
- z-GCshutdown.txt dentro de la carpeta de un experimento indica que Google Cloud asesinó a su virtual machine mientras estaba corriendo ese experimento. Si la virtual machine tuvo una muerte abruptamente violenta, este archivo no llega a grabarse,

Si una etapa del workflow aborta, entonces se detiene el avance del mismo. Usted podra ver con los nombres de carpeta hasta donde llegó.

En cada carpeta de un experimento, hay al menos un soft link, que es una carpeta virtuale que apunta a inputs que necesita el proceso para poder correr.

En cada carpeta de un experimento hay una copia del script .r que se corrió para ese experimento.

Sección Complementaria

Corra el script ~/dm2024a/src/workflow01/z507_workflow_base.r y analice los experimentos nuevos que se crean, y observe en todo su esplendor un Directed Acyclic Graph, una raíz común hasta FErf-0001 y luego nuevos experimentos TS-0002 HT-0002 y ZZ-0002