Desafío Analítica Avanzada

Descubriendo Exoplanetas

La misión espacial Kepler ha buscado exoplanetas (planetas fuera de nuestro sistema solar) durante la última década. Los candidatos a exoplanetas que se analizan, se corroboran mediante distintos procesos y luego de ellos, son registados una base de planetas que contienen una etiqueta. Al tener los datos con esta estructura y etiqueta, se vuelve posible usar técnicas de Machine Learning para automatizar el proceso de corroboración.

Su objetivo en éste desafío es construir un modelo de Machine Learning que detecte los planetas confirmados. Para ello considere los datos que se encuentran en el siguiente link:

`https://drive.google.com/file/d/1Ui9lIx8LeKaV6UuXT5VNbyz6XplGO2N5/view`

correspondientes a datos públicos sobre exoplanetas publicados por el *California Institute of Technology (CalTech)* y la *National Aeronautics and Space Administration (NASA)* de los Estados Unidos.

No se espera que sea ni astronomo ni menos experto en exoplanetas, pero si es de interés puedes ingresar el siguiente link para ver la descripción de las variables entregadas en el archivo csv.

`https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html`

Como ya fue comentado antes, se espera generar un modelo de Machine Learning que pueda automatizar el proceso de identificar qué candidatos son exoplanetas confirmados y cuales no.

Antes de comenzar a desarrollar el desafío, **lea todo el documento primero (incluyendo las indicaciones al final de este).**

Identificación y Preparación de Datos

La primera tarea a realizar es una identificación de las variables, en particular de la variable objetivo, estudios estadísticos y manipulación, como se indica a continuación:

1. Identifique la variable objetivo para modelar
2. Identifique variables que presentan un valor promedio muy distinto entre planetas confirmados y falsos positivos. Comente
3. Eligiendo una técnica de su preferencia, reduzca la cantidad de variables a la mitad procurando causar el menor impacto posible en la capacidad predictiva.
4. Identifique los tipos de variables que contiene el set de datos y ejecute–si corresponde–algún tipo de codificación o estandarización en las variables.

Entrenamiento y Optimización del Modelo

Luego de completar la primera parte de inspección y preparación de datos, ahora estamos listos para empezar con la modelación. En esta parte se espera que realices lo siguiente:

5. Construir set de datos que nos permitan entrenar y testear nuestro modelo
6. Entrene un modelo de Machine Learning a su elección y obtenga métricas para medir su desempeño
7. Utilice una técnica de su preferencia para hacer una optimización–si corresponde–de los hiperparámetros de su modelo
8. Re-entrene su modelo con la combinación de hiperparámetros encontrada y compare bajo las mismas métricas utilizadas antes con el modelo inicial (dado que estamos optimizando el modelo, se espera que este mejore su desempeño)

Evaluación del Modelo

Ya tenemos nuestro modelo entrenado y optimizado en la sección anterior, por lo que ahora realizaras algunos pasos que sirven para la evaluación del modelo:

9. Calcule las métricas que son apropiadas para el modelo entrenado
10. Si nuestro modelo nos indica que un candidato es confirmado como planeta, ¿Qué tan seguros podemos estar de que éste sea efectivamente un planeta? Comente
11. Construya la matriz de confusión para visualizar el comportamiento de su modelo.

¡Y listo! Eso es todo el desafío. Esperamos que puedas completar con éxito todas las tareas propuestas.

Algunas indicaciones

Aquí te dejamos algunas indicaciones para poder completar con éxito el desafío

a) Se espera que este desafío se realice en un **tiempo de no más allá de 4 horas** (ojo que este tiempo es de referencia)

b) La forma de solucionar el desafío nos dará una guía de tu manera de trabajar y también tus conocimiento. Si puedes o no resolver todo el desafío, no es necesariamente lo más importante. En este sentido, por favor envía lo que puedas hacer sin temor :).

c) Las respuestas deben venir en un jupyter notebook, R notebook, R Markdown o script de R. En este último caso, el código debe venir comentado para un correcto entendimiento.

d) Tienes hasta el **Lunes 14 de Octubre**, hasta las **23:59 horas** para entregar el desafío. La entrega se hace mediante correo electrónico a bgalasso@embonor.cl.

¡Todo el éxito!