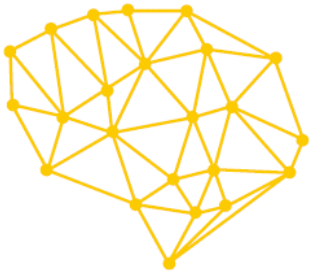

UNIVERSITYHACK 2020®
DATATHON

MINSAIT LAND CLASSIFICATION

COMENTARIOS

EQUIPO ASTRALARIA

1. Introducción

Para realizar la selección de los equipos finalistas del reto Minsait Land Classification de esta edición de la UniversityHack, inicialmente se obtienen las métricas a partir del fichero de predicciones presentado, ordenándolas de mayor a menor.

Vuestro equipo está por encima del percentil 90%, calculado considerando un rango de 0 a la métrica del 1º. 12 equipos cumplían esta condición.
La métrica obtenida ha sido 0,7006 (6º puesto).

Para obtener la ordenación final, el Jurado Nacional valora, además de la métrica obtenida (que es fundamental), distintos aspectos del código.

La valoración de los distintos jurados se agrega mediante el método Borda, estableciendo en la clasificación definitiva en el 4º puesto de un total de 21.

2. Comentarios sobre el proyecto

La opinión general del Jurado sobre el **análisis exploratorio** que realizáis es positiva, y como hecho destacable, es uno de los pocos equipos que se centran en la asimetría entre el dataset de modelar y estimar. También muy acertada la aplicación de PCA y t-SNE para fundamentar estrategias de data featuring.

La parte de la **creación de variables** está valorada en su conjunto de forma positiva sin destacar especialmente. Probablemente aún podría haber sido mejor si se hubieran añadido variables externas al dataset vinculadas a las coordenadas (algunos equipos han sido intensivos en esto). Otra aplicación bastante exitosa pero minoritaria entre los equipos ha sido la transformación de las variables GEOM mediante operaciones aritméticas. La aplicación de knn sobre las coordenadas está bien llevada, pero podría haberse dado un paso más generando nuevas variables a partir de distintos tipos de

distancias al centroide, por ejemplo. Las operaciones sobre los canales de color eran una buena idea, aunque era bastante difícil obtener valor de esas variables.

Desbalanceo. La valoración del jurado es muy buena en este punto, ya que no os limitáis a aplicar soluciones sencillas. No solo habéis aplicado un planteamiento amplio en cuanto a técnicas probadas, sino que también aplicáis iteraciones al random undersampling para evitar sesgos en la selección de registros. Como comentario, algún equipo ha usado AutoEncoder para generar registros de las clases minoritarias y técnicas bayesianas para optimizar el random undersampling. Este apartado es de los mejores de toda la competición.

En la parte de selección de **modelos** se valora positivamente el ensamblaje vinculado al desbalanceo y la aplicación de 2 algoritmos (seleccionados de una lista mayor) optimizados con grid search. Algunos equipos han optado (con éxito) por optimizaciones bayesianas, pero en general este es un punto que tenéis bien resuelto.

El proyecto está muy bien estructurado, con el código comentado de forma que facilita la comprensión y el seguimiento de lo realizado.

Esperamos que lo hayáis pasado bien con el reto.

Saludos
El Equipo de Cajamar UniversityHack