A continuación, comento los experimentos que hice para ver cómo variaba la curva de ganancia para diferentes valores de hiperparámetros. Adjunto imágenes que acompañan esto.

Cuando maxdepth es 20 y las hojas son pequeñas, la diferencia entre la curva de ganancia en train y test es notable a simple vista, hay un gran gap entre ellas, siendo la ganancia en train bastante superior que la de test. Por ejemplo, para un minsplit de 50 y un minbucket de 25, este gap es de $70.974.000. Esta diferencia tan grande se debe a que los arboles con profundidad alta y nodos chicos tienden al overfitting.

Pero a medida que vamos agrando las hojas, estas diferencias se achican. Cuando tenemos hojas extremadamente grandes, las curvan tienden a ser iguales. Probando con un minsplit mayor a 2000 podemos ver esto. Muestro tres imágenes de curvas de ganancia cuando las hojas son chicas, medianas y grandes.

[maxdepth_20_minsplit_50.png](/user_uploads/24/0/vDuWbiMD1XTdnJRGqZePGbwU/maxdepth_20_minsplit_50.png)

[maxdepth_20_minsplit_500.png](/user_uploads/24/d/Wx2rJqFLn6SXYCb65mulewKY/maxdepth_20_minsplit_500.png)

[maxdepth_20_minsplit_2800.png](/user_uploads/24/40/FxP8CleyMQAWk8B0dBCKRJf_/maxdepth_20_minsplit_2800.png)

Por otro lado, cuando maxdepth es 3, no existe demasiado gap entre las curvas de ganancia en train y test, sin importar mucho si los nodos son chicos o grandes. Tal vez la mayor ganancia la podemos notar en las posiciones 5mil a 10mil con minsplit menor a 1500, pero a medida que los agrandamos las curvas tienden a ser iguales. Para el caso de un minsplit de 50 y un minbucket de 25 este gap es de -4.068.000, es decir, la ganancia en test es mayor. Uno podría pensar que es un caso de underfitting debido a la simplicidad del modelo.

[maxdepth_3_minsplit_50.png](/user_uploads/24/e/23jAkSKRRRRm_ec6IXuiKvdu/maxdepth_3_minsplit_50.png)

[maxdepth_3_minsplit_500.png](/user_uploads/24/5a/USMR-RmPsiPvxSh18SzTyygm/maxdepth_3_minsplit_500.png)

[maxdepth_3_minsplit_3000.png](/user_uploads/24/6/nJU1rzlrDMml6bKYdcQKDh5d/maxdepth_3_minsplit_3000.png)

Por último, me quedó la duda de cómo sería la curva de ganancia para el caso de Elizabeth, para un maxdepth=6, minsplit=600 y minbucket=150, asi que ahí va:

[maxdepth_6_minbucket_150.png](/user_uploads/24/17/QruWI62ZvEGuWSOAO543anOh/maxdepth_6_minbucket_150.png)

Vemos que el gap entre las ganancias máximas es muy pequeño ($3.444.000). Las curvas se aproximan bastante.

Jugar con estos hiperparámetros complementa el ejercicio que hicimos en la clase 03, donde nos dimos cuenta qué hiperparámetros son los que realmente importan para el modelo, y los que no tenían sentido aplicar. Como conclusión, es importante destacar la importancia que tiene la buena selección de hiperparámetros de nuestro modelo, en este caso para el algoritmo cart ya que esto nos ayudará a evitar el underfitting y overfitting.