¿A mayor cantidad de arbolitos, mejora el modelo?

Según la teoría, generalmente al agregar más árboles a un ensemble puede mejorar la capacidad de generalización del modelo. Sin embargo, hay un punto de corte después del cual los beneficios obtenidos se vuelven muy chicos o incluso puede comenzar a disminuir un poco.

En mi caso, a partir de un ensemble de 50 árboles, la ganancia tiende a estabilizarse. Adjunto mi tabla y su respectivo gráfico.

¿La combinación de hiperparámetros que genera el mejor árbol unitario (arbolitos==1), es la misma que genera el mejor ensemble de 500 árboles?

No es la misma. La combinación de hiperparámetros que genera el mejor árbol unitario es { maxdepth=10, minsplit=1500, minbucket=20 } y la combinación que genera el mejor ensemble de 500 árboles es { maxdepth=10, minsplit=50, minbucket=10 } . Como mencioné anteriormente, al aumentar el número de árboles nos vamos acercando al mejor punto de corte o de rendimiento.

Cada set de hiperparámetros se ha corrido por DOS alumnos, cada uno con su semilla. ¿Qué sucede con la variabilidad de los resultados a medida que aumenta la cantidad de arbolitos?

Aún no están los datos. Cuando estén respondo la pregunta!

¿Qué otra conclusión puede obtener de los datos?

Por lo que pude observar en la planilla a nivel general, a partir de un ensemble de 5 árboles de decisión se puede notar las bondades que ofrece (ganancias mayores en kaggle en relación a un solo árbol). También se puede ver que al aumentar el número de árboles las ganancias tienden a ser parecidas (score en kaggle mayor a 50). Las que no llegan a este valor tienen en común un maxdepth y un minbucket muy pequeño. Aún la planilla no está completa, por lo que mis conclusiones pueden estar erróneas.