Árbol de decisión

Sebastian Makkos, Rodolfo Albornoz, Valeria Brzoza

Organización de datos, 1C23

Para este checkpoint tuvimos que realizar árboles de decisión. Lo primero que tuvimos que hacer fue la división del dataset disponible en “entrenamiento” y “testeo”, donde decidimos hacer la división en 70-30. Lo siguiente fue optimizar los hiperparametros con k-fold cross validation. Una vez realizado empezamos a generar y entrenar árboles. Utilizamos tanto el criterio Gini como Entropy y ambos fueron probados tanto sin podar, como podando.

Finalmente, nos quedamos con el árbol de entropy con poda, ya que fue el que mejor nos dio de los 4, según la métrica f1 score.

Con este árbol realizamos los demás estudios. Lo graficamos y vimos que sus atributos con mayor importancia eran el tipo de depósito y el agente. Pudimos observar que las métricas de entrenamiento y testeo no tuvieron gran diferencia. Por otro lado, la regla inicial está basada en el tipo de depósito.

- Si no hay depósito:
  Analiza las cancelaciones previas
  Si hay cancelaciones previas continuará por buscar las reservas previas no canceladas.
  Si no hay cancelaciones previas seguirá decidiendo en base al lead time

- Si hubo un depósito (Non Refund o Refundable):
  Analiza el país y dependiendo cual sea, podrá seguir por el customer_type o el agent

En conclusión, podemos decir que pudimos entrenar al mejor árbol, con criterio entropy, con un máximo de profundidad de 14, y notamos que era más conveniente para este criterio, utilizar poda al modelo. Al predecir el conjunto test sobre el modelo entrenado, obtuvimos un valor en 0.8416258287690978, con la métrica f1 score.