Ensambles

Sebastian Makkos, Rodolfo Albornoz, Valeria Brzoza

Organización de datos, 1C23

Para este checkpoint, tuvimos que realizar distintos ensambles. Creamos un clasificador KNN, donde optimizamos los hiperparametros con la métrica euclidean y con una cantidad de 6 vecinos. A la hora de sacar la predicción, obtuvimos un valor de 0.7851903194131578 para la métrica f1_score. El problema que podíamos haber encontrado, es que se overfitted el modelo en train.

Posteriormente, realizamos un clasificador SVM. En ella, probamos con dos formas de estandarizaciones, la normal y con min y max. También variamos los kernels, en lineal, polinómica y radial, y también realizamos un SVM con reducción de dimensionalidad. Comprobamos que el mejor clasificador era aquel que se realizó con reducción de dimensionalidad, ya que es aquel que se obtiene uno de los valores más altos para la métrica f1 score, 0.8002433090024329 y consideramos que no está overfitted.

Para el clasificador Random forest, podemos decir que a partir de las métricas analizadas para ambos conjuntos, el modelo podría estar overfitted en train. También analizamos que los atributos con mayor importancia son lead_time, country, y deposit_type y que es uno de los ensambles con valor más alto en la métrica f1_score, 0.8685073647802491.

Para el clasificador XGBoost, utilizamos 8 num_boost_round, un learning_rate de 0.1, una cantidad 150 estimadores. Y para la métrica del f1-score, obtuvimos un valor de 0.870248709526044, como el más alto de todos los ensambles

Para el ensamble híbrido tipo voting, le agregamos distintos tipos de clasificadores, como KNN, XGB, Logistic, y RF. Vimos que la diferencia del f1-score entre ambos conjuntos no es grande y por eso no la consideramos overfitted. Para la métrica f1_score, obtuvimos 0.8466327108092813.

Por último, para el ensamble híbrido tipo stacking, definimos los modelos base que tendrá el ensamble, el metamodelo de tipo logistic regression, consideramos al modelo overfitted y además, obtuvimos para la métrica f1 score un valor de 0.8698097403374417, como el segundo más grande entre todos.

Al analizar entre todos los ensambles modelados y entrenados, podemos concluir que a nuestra consideración, los mejores fueron XGBoost y Ensamble híbrido tipo voting, ya que por mas que están entre los 4 con mayor valor de métrica f1_score, eran los dos que no consideramos como overfitted.