KAGGLE COMPETITION

OBJETIVO

En esta competición hay que predecir el precio de un diamante en función de sus características.
Datasets:

- train.csv: contiene los precios para entrenar el modelo
- predict.csv: sin precios para realizar la predicción

Steps

I. Data Cleaning & Model Exploration

- Eliminar la columna de ID
- Cut, color and clarity: Convertir a valores numéricos (Label Encoder vs Diccionario
- Explorar primeros modelos con sklearn y elegir los mejores en función del RMSE y R2

II. Extra Trees Regressor

- ExtraTreesRegressor(n_estimators=1800, min_samples_split=10, max_depth= 50)
- RMSE: 538.47
- R2: 0.981557

III. Hist Gradient Boosting Regressor

- HistGradientBoostingRegressor(loss='least_squares', max_depth=150, min_samples_leaf=2)
- RMSE: 544.39
- R2: 0.980244

IV. Random Forest Regressor

- RandomForestRegressor(max_depth=10, max_features=10, min_samples_leaf=20, min_samples_split=15, n_estimators=50)
- RMSE: 629.77
- R2: 0.973514

V. H2O

- H2OAutoML(max_runtime_secs=1000, sort_metric='RMSE')
- Model: StackedEnsemble (RMSE: 533.132)
- Model: GBM_3 (RMSE: 539.697)

Conclusiones

El modelo ganador es el generado de manera automática por H2O:

StackedEnsemble

- RMSE: 533.132