Introduction to Machine Learning: 6th lesson – Random Forests

Introduction:

Decision trees leave you with a difficult decision. A deep tree with lots of leaves will overfit because each prediction is coming from historical data from only the few houses at its leaf. But a shallow tree with few leaves will perform poorly because it fails to capture as many distinctions in the raw data. Even today's most sophisticated modeling techniques face this tension between underfitting and overfitting. But, many models have clever ideas that can lead to better performance. We'll look at the **random forest** as an example. The random forest uses many trees, and it makes a prediction by averaging the predictions of each component tree. It generally has much better predictive accuracy than a single decision tree and it works well with default parameters. If you keep modeling, you can learn more models with even better performance, but many of those are sensitive to getting the right parameters.

Example:

You've already seen the code to load the data a few times. At the end of data-loading, we have the following variables:

* train\_X
* val\_X
* train\_y
* val\_y

import pandas as pd

*# Load data*

melbourne\_file\_path = '../input/melbourne-housing-snapshot/melb\_data.csv'

melbourne\_data = pd.read\_csv(melbourne\_file\_path)

*# Filter rows with missing values*

melbourne\_data = melbourne\_data.dropna(axis=0)

*# Choose target and features*

y = melbourne\_data.Price

melbourne\_features = ['Rooms', 'Bathroom', 'Landsize', 'BuildingArea',

'YearBuilt', 'Lattitude', 'Longtitude']

X = melbourne\_data[melbourne\_features]

from sklearn.model\_selection import train\_test\_split

*# split data into training and validation data, for both features and target*

*# The split is based on a random number generator. Supplying a numeric value to*

*# the random\_state argument guarantees we get the same split every time we*

*# run this script.*

train\_X, val\_X, train\_y, val\_y = train\_test\_split(X, y,random\_state = 0)

We build a random forest model similarly to how we built a decision tree in scikit-learn - this time using the RandomForestRegressor class instead of DecisionTreeRegressor.

from sklearn.ensemble import RandomForestRegressor

from sklearn.metrics import mean\_absolute\_error

forest\_model = RandomForestRegressor(random\_state=1)

forest\_model.fit(train\_X, train\_y)

melb\_preds = forest\_model.predict(val\_X)

print(mean\_absolute\_error(val\_y, melb\_preds))

191669.7536453626

Conclusion:

There is likely room for further improvement, but this is a big improvement over the best decision tree error of 250,000. There are parameters which allow you to change the performance of the Random Forest much as we changed the maximum depth of the single decision tree. But one of the best features of Random Forest models is that they generally work reasonably even without this tuning.