Data Analysis and Machine-Learning

Chapter 5.3:

Feature Selection (3)

Variable Weight Evaluation Methodologies



1. Evaluation of Variable Weights

There are several ways of evaluating the weights of variables to decide which independent variables to select and discard. The most representative one includes the beta coefficient, as we have covered in chapters 2 and 3. One limitation of using beta coefficients is that despite their accuracy and intuitiveness, beta coefficients are only applicable to limited types of models including linear models and generalized linear models. In this context, the weight evaluation methods introduced in this chapter such as the permutation importance method, eli5, and shap, have relative advantages in terms of their wide applicability, as they are also applicable to various other algorithms other than linear models.

2. Eli5

Eli5 provides visualization regarding the effects of each variable. The comparative advantage of eli5 is that weights of the variables are visible for various other algorithms other than linear models.

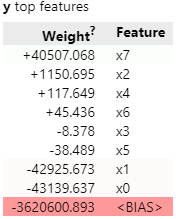
!pip install eli5

import eli5

base\_model = LinearRegression().fit(features, target)

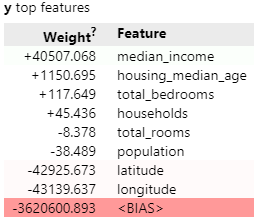
eli5.show\_weights(base\_model)

output:



#Import feature names

eli5.show\_weights(base\_model, feature\_names = features.columns.tolist())



Accordingly to the eli5 evaluation, it can be stated that all 8 variables are concluded to be significant in terms of weights. Inter alia, longitude, latitude, median income, and housing median age variables are concluded to be particularly significant in terms of their weight impacts on the target variable.

3. Permutation Importance

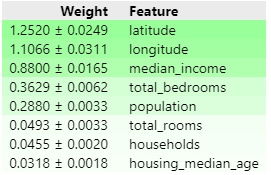
The peculiar feature of permutation importance evaluation method is that it calculates weight scores by randomly shuffling the variables.

from eli5.sklearn import PermutationImportance

base\_model = LinearRegression().fit(features, target)

perm = PermutationImportance(base\_model).fit(features, target)

eli5.show\_weights(perm, feature\_names = features.columns.tolist())



Features are aligned in order of weights. Accordingly to the permutation importance evaluation, it can be implied that latitude, longitude, median income are particularly significant in terms of weight impacts on the target variables (\*total bedrooms and population variables are also includable, depending on the analyst’s interpretation).

Comprehensively speaking based on the evaluations with eli5 and permutation importance, it can be concluded that the three variables of longitude, latitude, median income are the main elements that explain the target variable.

4. Shap

To be sure with the result, it would be better to examine the variable weights in various perspectives. In this context, Shap values may provide additional insights regarding feature selection.

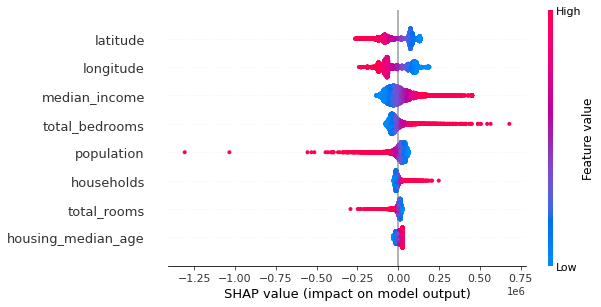
import shap

base\_model = LinearRegression().fit(features, target)

explainer = shap.LinearExplainer(base\_model, features)

shap\_value = explainer.shap\_values(features)

shap.summary\_plot(shap\_value, features)



Analogous to the results from eli5 and permutation importance, longitude, latitude, and median income can be concluded as the most explanatory variables with significant weights.

Note that, according to the evaluation result, total rooms and total bedrooms are concluded to have opposite effects on the median house value, which does not really make sense intuitively. It can be inferred that such problem is most likely resulting from multicollinearity problem, which is why it is significant to comprehensively consider weight values with multicollinearity values (i.e., correlation, causality) that we covered in the earlier chapter. No one evaluation method can be stated as absolute. It is the analyst’s capacity to interpret the comprehend the evaluation result synthetically, and to decide which variables are most appropriate to adopt in the final model.