

Model Explainability

Ryan Miller

Introduction

- ▶ Model evaluation metrics including accuracy, AUC, and F1 provide an assessment of *how much* a model has learned
 - ▶ These metrics do not address *what* the model has learned
- ▶ Approaches to model explainability generally fall into two categories:
 - ▶ **Global approaches** help us understand what patterns the model has learned across the entire dataset
 - ▶ **Local approaches** explain why a particular prediction was made for a specific observation

Built-in Importance Measures

- ▶ Some modeling algorithms have built-in ways of measuring importance.
 - ▶ Tree-based algorithms, including random forests and xgboost, have gain-based metrics that quantify how much the model improves from splits involving a given variable
- ▶ Our focus this week will be on *model agnostic* approaches, or methods that can be used for *any* machine learning model

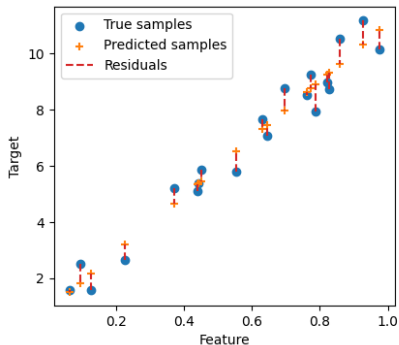
Permutation Feature Importance

- ▶ Permutation feature importance randomly shuffles the values of the target feature to break its relationship with the outcome variable
 - ▶ The modified data are given to the trained model, and the performance score is compared with the model's baseline score (without permutation)
 - ▶ Features that cause larger performance deterioration when permuted are considered more important
- ▶ To account for variability across random shuffles, this process is generally repeated several times for each variable

Permuting an Important Feature

Effect of permuting a predictive feature

Linear model without feature permutation
Mean Absolute Error: 0.51



Linear model with feature permutation
Mean Absolute Error: 2.28

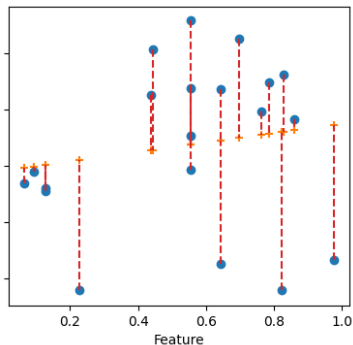


image source: https://scikit-learn.org/stable/modules/permutation_importance.html

Permuting an Unimportant Feature

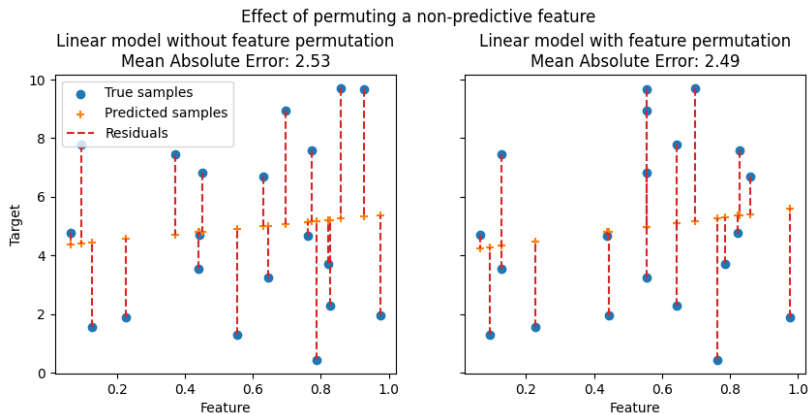
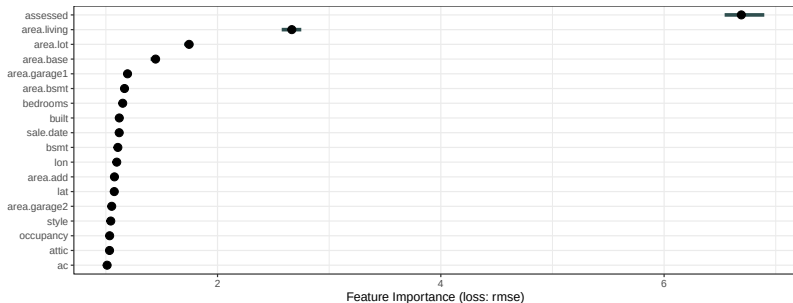


image source: https://scikit-learn.org/stable/modules/permutation_importance.html

Example - Home Price Prediction



Limitations of Permutation Importance

When features are highly collinear there is little change in model performance when a feature is permuted because most of the same predictive information can be found in a correlated feature. A few potential strategies to deal with this:

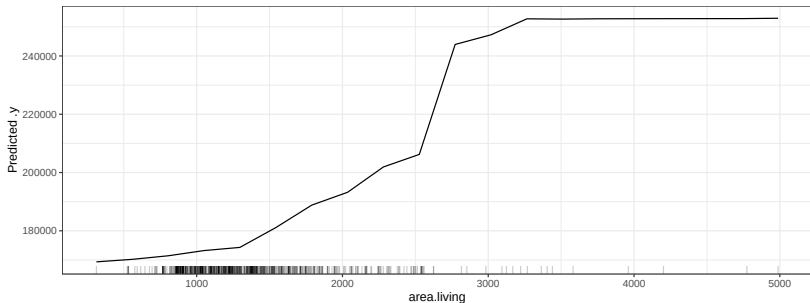
1. Apply PCA as a pre-processing step. Principal components are inherently independent.
2. Identify groups of correlated features by performing clustering on the correlation matrix, then permute entire groups of features instead of single features.

The `sklearn` documentation has a nice example of the latter approach at [this link](#).

Partial Dependence Plots

- ▶ Permutation feature importance can help determine *which* features have the greatest global importance, but it doesn't explain *how* these features impact predictions
- ▶ A partial dependence plot (PDP) shows the marginal effect of a feature on the average predicted outcome
 - ▶ This is achieved by fixing the predictor at a sequence of values, leaving all other features unchanged, and averaging the model's resulting predictions
 - ▶ The resulting curve describes how the model's average prediction changes as the feature varies.

Example - Home Price Prediction



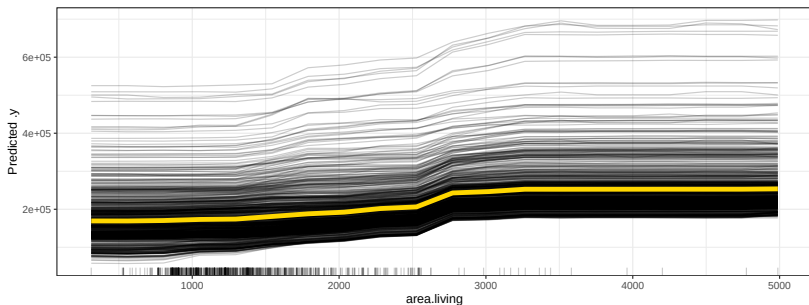
Limitations of PDPs

- ▶ PDPs require manual interpretation, so for most data sets you'll only look at them for features that were already identified as important
- ▶ The average prediction might involve nonsensical combinations of features, such as homes with living areas that are smaller than their basement areas
 - ▶ In this sense, PDPs also work best when features are independent and can be misleading when features are highly collinear
- ▶ Finally, PDPs can hide heterogeneous effects (interactions) because they are averaging across the entire dataset

Individual Conditional Expectation Plots

- ▶ Two observations can have very different responses to changes in a predictor, even when the PDP suggests only a modest average effect
- ▶ Individual Conditional Expectation (ICE) plots show the relationship between a feature and the predicted outcome across the entire data set, plotting one line per sample
 - ▶ ICE plots can reveal interactions that might not be apparent in a PDP due to averaging

Example - Home Price Prediction



The average marginal effect is shown in yellow. If all the ICE curves were roughly parallel, we could conclude the effect of living area is similar across homes. Here there is mild heterogeneity of the effect, evidenced by curves that cross.

Local Explainability

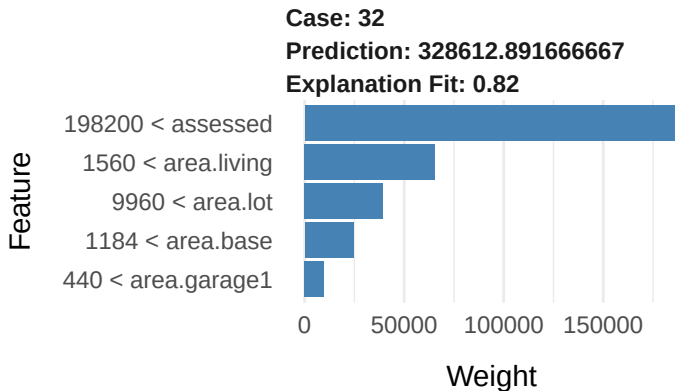
- ▶ *Global explainability* approaches answer questions like:
 - ▶ “What role does living area play in the model?”
- ▶ In contrast, *local explainability* addresses questions like:
 - ▶ “Why did the home sold on 4/7/2005 sell for \$325000?”

Local Interpretable Model-Agnostic Explanations (LIME)

- ▶ LIME approaches local interpretability by generating hypothetical data points near the target observation
- ▶ A simple model (typically weighted linear/logistic regression) is fit to this set of hypothetical local data points
 - ▶ Coefficients from this model are then used to explain the model's prediction for original target observation

Example - Home Sales

Most LIME implementations will default to binning continuous predictors (discretization) to facilitate easier human interpretations and account for non-linearity



Limitations of LIME

- ▶ Local explanations are also not guaranteed to reliably reflect global model behavior, so inferences are limited to the sample being explained
- ▶ Additionally, LIME is dependent on the perturbation strategy used to construct local neighborhoods
 - ▶ Thus, local approximations can be unstable, and different neighborhoods can produce different explanations
- ▶ Because LIME explanations depend on how the local neighborhood is generated, researchers developed Shapley Additive Explanations (SHAP), which provides explanations grounded in cooperative game theory

Shapley Additive Explanations (SHAP)

- ▶ SHAP approaches seek to fairly distribute credit for a prediction across the input features
 - ▶ More specifically, individual predictions are decomposed into a baseline prediction (ϕ_0) and the additive contributions of each feature (ϕ_j for $j \in \{1, \dots, p\}$):

$$\hat{y}_i = \phi_0 + \phi_1 + \phi_2 + \dots + \phi_p$$

For example, if a model predicts a certain home will sell for \$310,000 and the average sale price in the data set is \$180,000 there is a gap of \$130,000 that SHAP seeks to allocate across the features

Shapley Additive Explanations (SHAP)

- ▶ SHAP calculates the marginal contribution of each feature by evaluating how the model's prediction changes for coalitions that do and do not contain that feature
- ▶ Consider a simple scenario with only two features, assessed and area.living.
 - ▶ Shown below are all possible coalitions and some hypothetical predictions using those coalitions:

Features Available to Model	Prediction
{ } (baseline)	\$180,000
{Living Area}	\$240,000
{Assessed}	\$220,000
{Living Area, Assessed}	\$310,000

Shapley Additive Explanations (SHAP)

To find the SHAP value for assessed, we average its marginal contributions across each coalition it can be added to:

Initial Coalition	Prediction Before	Marginal Contribution of Assessed
{}	\$180,000	$\$220,000 - \$180,000 = \$40,000$
{Living Area}	\$240,000	$\$310,000 - \$240,000 = \$70,000$

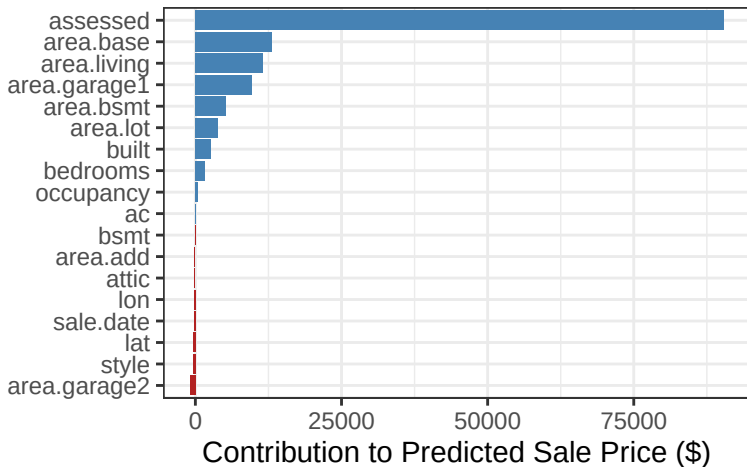
$$\text{So, } \phi_{\text{Assessed}} = \frac{40,000 + 70,000}{2} = 55,000$$

SHAP Details

- ▶ When evaluating a coalition, SHAP uses the trained model and averages predictions over the features excluded from the coalition
 - ▶ Thus, a feature being absent in a coalition does not mean the feature is set to zero or omitted from the model. Instead, SHAP treats the feature as unknown and averages over plausible values from the data
- ▶ The basic idea, using the {Living Area} coalition as an example, is to answer a question like: “What would the model predict if we knew Living Area = 2500 but had no information about Assessed?”

Example - Home Sales

SHAP values for the home sold on 4/7/:



Using SHAP for Global Importance

- ▶ Although SHAP was originally developed for local explanations, it can also be used to assess global importance.
- ▶ A common approach is to calculate the average absolute SHAP value for each feature
 - ▶ Features with a larger average absolute value tend to have a larger impact on model predictions
- ▶ A strength of using SHAP to assess global importance is the way in which it handles correlated features.

Key Takeaways

Rather than treating these methods of local and global explainability as distinct methods to choose between, they can be used to answer a series of progressive questions:

- ▶ Which variables matter? Answer using Permutation Importance
- ▶ How do they affect predictions on average? Answer using PDP
- ▶ Does the effect vary across observations? Answer using ICE
- ▶ Why was this specific prediction made? Answer using LIME or SHAP