Introduction

Permutation Feature Importance (PFI) is believed to be a model-agnostic interpretability method since it is not based on the internal components of the predictive model. A method called PIMP algorithm, (Altmann et al. 2010) was suggested to correct the biases of classic measures of importance, and proved to apply to more than just random forests, but also to any learning algorithm that gives feature relevance scores, such as linear models, logistic regression and support vector machines. This approach can provide more appropriate feature ranking by attributing significance P-values by permuting the results and provide improved interpretability among various models. [1]

Advantages

1. Model-agnostic
   It can be applied to any existing pre-trained model (SVM, RF, XGBoost, neural network, etc.) without being able to access the internal weights or structure of the model. [2]
2. Intuitive and easy to explain
   The method evaluates the significance by monitoring the drop in performance with the perturbation of a feature. The outcomes are strongly linked to the actual business metrics (F1 and AUC-PR), which is why it is simple to communicate to non-technical staff. [2]
3. Quantifiable
   You are allowed to select any evaluation metric of interest and therefore the importance is the performance change that you really care about.
4. Can detect the effects of nonlinearities and interactions
   When it is possible to model these relationships using black-box models, shuffling will expose how much the feature is actually used in the model. [3]
5. Can be used for feature selection decisions
   It may be applied to screen out features that do not contribute to the metric, and as the foundation of downstream feature selection or dimensionality reduction. [1]

Limitations

1. Computational cost
   Permutation based importance estimation methods have importance based on repeatedly permuting each feature and reassessing the score of the model. This is computationally expensive, especially when the number of features is large or the

prediction time of the model is slow. [2]

2. Correlation issues
   In cases where two features are very correlated or a set of features are very correlated, the permutation of one of them usually would cause a very little change in the score of the model. Consequently, the actual contribution of that feature could be underrated. [4]
3. Metric sensitivity
   The significance which is calculated through the permutation feature importance varies with the measure of evaluation you use. The various metrics (e.g. F1, AUC-PR, or accuracy) highlight different error types, which results in different feature rankings. [2]

Code

```
>>> from sklearn.datasets import load_diabetes
>>> from sklearn.model_selection import train_test_split
>>> from sklearn.linear_model import Ridge
>>> diabetes = load_diabetes()
>>> X_train, X_val, y_train, y_val = train_test_split(
...     diabetes.data, diabetes.target, random_state=0)
...
>>> model = Ridge(alpha=1e-2).fit(X_train, y_train)
>>> model.score(X_val, y_val)
0.356...
```

```
>>> from sklearn.inspection import permutation_importance
>>> r = permutation_importance(model, X_val, y_val,
...                            n_repeats=30,
...                            random_state=0)
...
>>> for i in r.importances_mean.argsort()[::-1]:
...     if r.importances_mean[i] - 2 * r.importances_std[i] > 0:
...         print(f"{diabetes.feature_names[i]:<8}"
...               f"{r.importances_mean[i]:.3f}"
...               f" +/- {r.importances_std[i]:.3f}")
...
s5      0.204 +/- 0.050
bmi     0.176 +/- 0.048
bp      0.088 +/- 0.033
sex     0.056 +/- 0.023
```

Summary

Permutation Feature Importance (PFI) is a powerful, model-free, interpretability approach. Its operation is as follows: the values of a feature are perturbed randomly and the performance of the model is monitored as the values are perturbed. The extent of degradation shows the influence that such a feature has on predictions, and black-box models become more understandable, and the most significant input variables are

identified to use them in making decisions and selecting features.PFI though is highly correlated and computationally expensive. In practice, it is recommended to use PFI together with other interpretability methods to gain a more comprehensive and reliable insight into model behavior.

[1] Altmann, A, Toloşi, L, Sander, O & Lengauer, T 2010, 'Permutation importance: a corrected feature importance measure', *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347.

[2] *5.2. Permutation feature importance* 2025, scikit-learn.

[3] Breiman, L 2001, 'Random Forests', *Machine Learning*, vol. 45, no. 1, pp. 5–32.

[4] *Permutation Importance with Multicollinear or Correlated Features* 2025, scikit-learn, <https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance_multicollinear.html?utm_source=chatgpt.com>.