Feature importance introduction

Sijin Zhang

There are mainly two types of feature importance estimation methods

Gini importance

Permutation importance

There are mainly two types of feature importance estimation methods

Gini importance

Can be only used for random forest

Permutation importance

Can be used by any types of model

There are mainly two types of feature importance estimation methods

Gini importance

Can be only used for random forest

Less intuitive, based on Gini impurity decrease

Permutation importance

Can be used by any types of model

More intuitive, capable of using any metrics such as POD, ACC etc.

There are mainly two types of feature importance estimation methods

Gini importance

Can be only used for random forest

Less intuitive, based on Gini impurity decrease

Not that easy to understand

Permutation importance

Can be used by any types of model

More intuitive, capable of using any metrics such as POD, ACC etc.

Very easy to understand

There are mainly two types of feature importance estimation methods

Gini importance	Permutation importance
Can be only used for random forest	Can be used by any types of model
Less intuitive, based on Gini impurity decrease	More intuitive, capable of using any metrics such as POD, ACC etc.
Not that easy to understand	Very easy to understand
N/A	Can only be used for one particular trained model

There are mainly two types of feature importance estimation methods

Gini importance	Permutation importance
Can be only used for random forest	Can be used by any types of model
Less intuitive, based on Gini impurity decrease	More intuitive, capable of using any metrics such as POD, ACC etc.
Not that easy to understand	Very easy to understand
N/A	Can only be used for one particular trained model
If RF is overfitting, then the importance does not mean much	Less chance of “overfitting” since it uses independent dataset
Require less dataset	Require more dataset (e.g., need to split and shuffle dataset)
Very cheap to run	Relatively more expensive

There are mainly two types of feature importance estimation methods

Gini importance	Permutation importance
Can be only used for random forest	Can be used by any types of model
Less intuitive, based on Gini impurity decrease	More intuitive, capable of using any metrics such as POD, ACC etc.
Not that easy to understand	Very easy to understand
N/A	Can only be used for one particular trained model
If RF is overfitting, then the importance does not mean much	Less chance of “overfitting” since it uses independent dataset
Require less dataset	Require more dataset (e.g., need to split and shuffle dataset)
Very cheap to run	Relatively more expensive

Both methods are not very good when features are correlated (e.g., if two features are similar, then they may share the importance, so the importance for both features will be underestimated)