Overview of TidyModels

TidyModels is an ecosystem that groups several R packages for machine learning and statistical models, focusing on 3 main steps: preparation, fitting, and evaluation

Components of `tidymodels`

Preparing data

1. `rsample` This package supports initial partition of data, which is a common practice for building machine learning models. With `rsample`, we can easily split data into a training set (for estimating model parameters) and a test set (for evaluating model performance). `rsample` also supports resampling of data such as bootstrapping and cross-validation used for model tuning and selection.

2. `recipes` This package offers ways to pre-process and manipulate data before the actual modeling work. Depending on the nature of model, data scientists need to determine the pre-processing steps and the order of applying these steps. This can be handled by creating a recipe object containing outcome and predictors, followed by sequence of pre-processing steps (imputation, transformation, dummy variable creation, normalization...).

Fitting models

3. `parsnip` This package provides the infrastructures to specify and fit models in a consistent manner. Defining a model includes the model type (random forests or LDA...), mode (classification or regression) and computational engine indicating how the model is fitted. From here, we can train our data to obtain a parsnip model object that can be used directly for prediction.

4. `tune` This package helps optimize hyper-parameters of a model. In case some certain parameters require tuning, we can define the model and explicitly specify parameters required tuning. We then use a function from `tune` to compute performance metrics for the set of tuning parameters to select the best one for our model.

Evaluating models

5. `yardstick` This package aims to evaluate how well models perform. It contains a variety of functions to calculate the performance metrics including classification metrics (sensitivity, specificity...), regression metrics (R squared, rmse...) and curve function (ROC curve...)

Supported data and models

`tidymodels` works with different types of data including numerical data, categorical data, text data, time series data and spatial data, making it suitable to a wide range of applications. `tidymodels` currently supports many existing ML and statistical models such as linear regression, logistic regression, regression tree, random forest, SVM... Some models like LOESS or smoothing spline are not yet integrated into `tidymodels`. In those cases, we can use base R functions as alternatives.

Advantages of adopting `tidymodels`

Standardized model interface: `tidymodels` provides a standard syntax for specifying and fitting model. This allows users to focus on more important tasks instead of having to memorize the syntax of different models.

Consistent workflow: `tidymodels` consists of components that correspond to specific steps of the data modeling workflow. Following the similar workflow structure makes the code more logical and easy-to-understand, which enhances collaboration between team members.

Modular structure: `tidymodels` follows a modular design, in which the structure is partitioned into bite-sized chunks. This allows users to modify things like pre-processing steps or computational engine quickly and effortlessly without having to rewrite the entire workflow.

Integration with `tidyverse`: `Tidymodels` was built on some core `tidyverse` packages, thus allowing users to smoothly run through the whole data science project life cycle (from data manipulation and visualization to data modeling).

Disadvantages of using `tidymodels`

Learning Curve: To effectively use `tidymodels`, proficiency in working with tidyverse and R is required. Hence, it might take time for a complete beginner to grasp the whole concept of `tidymodels`.

Maturity: `tidymodels` is still a relatively new package compared to well-established alternatives like `caret`. As a result, it may not have reached the same level of maturity in terms of feature completeness, stability, and extensive community support.