STATS 7022 – Data Science PG Assignment 2

Dang Thinh Nguyen

2024-07-11

Question 4: Technical Communication

The `tidymodels` package in R is a collection of packages designed for modeling and statistical analysis using tidyverse principles (Kuhn & Wickham, 2024).

The core components of `tidymodels` package can be categorised into subsets according to analysis workflow (Wright et al., 2021), which include:

1. Data Preprocessing: `rsample` (splitting data into training and testing sets or resampling data) and `recipes` (preprocessing tasks like imputation, normalisation and feature engineering).
2. Model training and prediction: `parsnip` (selecting model, fitting data and predict) and `workflows` (combining pre-processing and modeling steps into a single workflow).
3. Model Evaluation: `yardstick` (using metrics like RMSE, accuracy, ROC AUC, etc.)
4. Model Optimisation: `tune` (hyperparameter tuning to improve performance)

The `tidymodels` package supports a wide range of data types and models:

1. Data Types:
   - Structured data (data frames and tibbles) is the most common type of data used in `tidymodels`. Ideal for tabular data where each column represents a variable and each row represents an observation.
   - Other types of data such as time series data or text data need to be preprocessed using additional packages (e.g. `tsibble` for time series data, `textrecipes` for text data) before feeding in a model with 'tidymodels' package.
2. Models:
   The `tidymodels` package can be applied to multiple machine learning models (Kuhn & Vaughan, n.d.). Some common types of both classical and modern machine learning algorithms supported by the package and their functions are:
   - Linear Models: Linear Regression (`linear_reg`) or Logistic Regression (`logistic_reg).
   - Tree-Based Models: Decision Trees (`decision_tree`), Random Forests (`rand_forest`) or Boosted Trees (`boost_tree`).
   - Support Vector Machines (SVM) (`svm_linear`, `svm_poly`, `svm_rbf`).
   - Nearest Neighbors: K-Nearest Neighbors (KNN) (`nearest_neighbor`).
   - Discriminant Analysis Model (`bayes_gml`): Linear Discriminant Analysis (LDA) (`discrim_linear`) or Quadratic Discriminant Analysis (QDA) (`discrim_quad`).
   - Neural Networks: Multilayer Perceptron (`mlp`).

Advantages of `tidymodel` that can be considered are:

1. Integration: the `tidyverse` package is a combination of multiple small packages that allow to facilitate data manipulation, visualisation, and analysis.
2. Consistency: A uniform API across different modeling packages ensures a smoother learning curve and consistent workflow.
3. Reproducibility: Built-in support for reproducible workflows through `workflows` and `rsample`.

The `tidymodels` package also have several disadvantages, which are:

1. Learning Curve: While `tidymodels` aims for consistency, the breadth of the ecosystem can be overwhelming for beginners.
2. Performance: For very large datasets or highly specialised models, the abstraction layer might introduce some performance overhead compared to using lower-level packages directly.

Adopting `tidymodels` can significantly streamline the data science workflow, offering a consistent and powerful suite of tools. While there is a learning curve and performance trade-offs, the benefits of integration, consistency, and reproducibility make it a strong choice for many data science projects.

Reference list

1. Kuhn, M & Wickham, H 2024, *tidymodels: easily install and load the 'Tidymodels' packages*, CRAN, viewed 9 July 2024, <https://cran.r-project.org/web/packages/tidymodels/index.html>.
2. Wright, C, Ellis, SE, Hicks, SC & Peng, RD 2021, *Tidyverse skills for data science*, The Johns Hopkins Data Science Lab, viewed 9 July 2024, <https://jhudatascience.org/tidyversecourse/model.html#summary-of-tidymodels>.
3. Kuhn, M & Vaughan, D n.d., *Function reference*, tidymodels, viewed 9 July 2024, <https://parsnip.tidymodels.org/reference/index.html#models>.