STATS 7022 - Data Science PG
Question 4: Technical Communication

Possakorn Kittipipatthanapong a1873765

Introduction

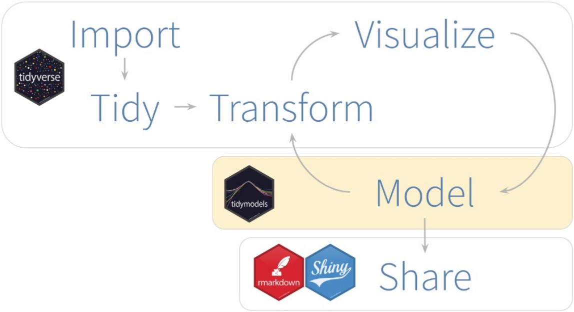

Figure 1: Overview Analysis tools in R

Tidymodels is a collection of R packages designed for comprehensive data analysis, with a particular emphasis on modeling and machine learning following in **Figure01**. Building on the foundation of Tidyverse's design principles, syntax, and data structures, Tidymodels introduces a streamlined workflow for managing different data types and models. This tool has distinct advantages and disadvantages, which we will explore in detail.

Package Components

The majority of a data scientist's workflow, as illustrated in **Figure02**, consists of data pre-processing, training, and validation stages. The Tidymodels Workflow aims to streamline these processes through several core components listed below:
01 rsample: This package offers a variety of data resampling methods which are essential for initial data splitting and cross-validation. These steps are crucial in model evaluation.
02 recipes: This package is used to define pre-processing steps for all potential variables. It includes scaling, normalization, handling missing data, and providing dummies for categorical variables.
03 parnip: This package is used to design the structural model specification with a variety of model types. It also provides the option to select the computational engine appropriate for different situations. When combined with the **tune** package, it allows for hyperparameter tuning to optimize the model.
04 yardstick: This package is used to measure the model performance with various types of metrics such as accuracy, R-squared (rsq), or root mean square error (rmse). By employing these components, the Tidymodels Workflow helps to make the data scientist's workflow more efficient.

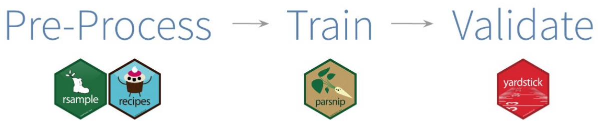

Figure 2: Data Analysis Workflow mapping with core packages in Tidymodels

Types of Data and Models Supported

Tidymodels works with various data types including numerical, categorical, factor, date-time and also string. It also supports a broad range of predictive model consisted of regression, classification, clustering. For example, Logistic regression, Random forests, K-nearest neighbors, decision trees, and SVM.

Advantages of Adopting Tidymodels:

01 Organized and Structured Interface: Tidymodels provides a consistent and unified syntax across various models, leading to enhanced code readability and improved reproducibility.
02 Flexibility and Extensibility: The framework allows users to customize various steps, including preprocessing, model specifications, and metric evaluations. This flexibility makes Tidymodels adaptable to a wide range of tasks and requirements.
03 Seamless Integration with Other Packages: For machine learning implementation in R, tidyverse is one of the most popular packages for data preparation and other analyses. Tidymodels offers a seamless ecosystem that simplifies data manipulation and visualization through model outputs, effectively integrating with the tidyverse package and others.

Potential Disadvantages:

01 Learning Curve: The syntax and principles of this package are pretty different from base R and other modeling packages, implying an initial learning curve.
02 Performance: While this package focuses on clarity and ease of use, it may face performance limitations when dealing with large datasets due to computational constraints.