

Machine learning (ML) and statistical algorithms have been significantly used in various applications, such as data classification, predictive regression, and feature selection. As the need for data-driven insights continues to grow, there is an increasing demand for exploratory and predictive data analysis to support business decision-making, academic research, and other applications. Identifying the best model that has optimal performance for a specific dataset usually consumes much time depending on the purpose of analysis. Although there are many packages that provide pre-built machine learning or statistical models, users still need time to load various suitable packages or functions, optimization of hyperparameters, validate the model, acknowledge the statistical relationship between each random or bivariate variable, and so on. This paper presents a package for R, “Quantlyzer” that contains various popular algorithms from machine learning and statistics. This tool aims to make automated data analysis more convenient for all different levels of users from no data analytics experience to domain experts to improve the efficiency of analyzing data. A workflow pipeline of exploratory analytics that contains various popular descriptive analysis techniques (e.g., Pearson Correlation Coefficient, a statistical summary of each variable, data visualization), statistical algorithms (e.g. Ridge regression, Ordinary Least Squares (OLS), Decision tree), machine learning models (e.g. Support Vector Machine (SVM), Random Forest, eXtreme Gradient Boosting (XGBoost), Gradient Boosting Machines (GBMs)), and Automated machine learning (AutoML) were ensembled in this package. The five-fold cross-validation is always used in every machine learning model to avoid overfitting or selection bias. A dataset with the index of soil organic carbon as the dependent variable and Near-Infrared Spectroscopy (NIRS) as independent variables was used to evaluate

the performance of the predictive regressions for the package. Another dataset, which is based on estimating different levels of dicamba damage on the soybean plot with extracted image features as independent variables were used to testify classification performance by using the Quantlyzer. The result shows that a pipeline using ensembled various machine learning and statistical algorithms can not only generate the report within 2 hours but also provide broader information including data visualization, statistical analysis, and machine learning results with a few lines of code. This study demonstrates a possible method to develop an automated data analysis platform with various techniques to help both academics and industry to discover patterns in data. The goal of this project is designed to enhance the data analysis efficiency for users by minimizing the need for manual code input, ultimately reducing the effort and time consumption. The source code is freely available through GitHub (<https://github.com/tianfengkai/quantlyzer>).