Introduction

There are a lot of predictive modeling algorithms in machine learning (ML) in engineering applications. Predictive modeling is the problem of developing a model using historical data to predict new data where we do not have the answer. Predictive modeling can be described as the mathematical problem of approximating a mapping function (f) from input variables (X) to output variables (y). This is called the problem of function approximation. Generally, we can divide all function approximation tasks into classification and regression tasks.

Regression models include Single and Multiple Linear Regression, Decision Tree, Polynomial, Random Forest, and Support Vector. Classification models include Decision Tree, K-Nearest Neighbors, Kernel SVM, Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machine.

One of the most frequently asked questions in the data science community seeks to determine which regression or classification model is best suited to be used on different datasets. This project aims to build a python package that will help you evaluate your regression models to select the best model for your dataset quickly and efficiently. Provide the file path to your dataset with some optional parameters and watch this package do the rest of the work for you.

User Profile

Targeted users for this package are machine learning engineers, data scientists, data analysts, and students. This tool also caters to people who leverage machine learning models without solid ML concepts.

Data Sources

I have provided some data for testing within the data folder in the project structure. The sales.csv and data.csv files are for the regression models testing, whiles the data_classification.csv is for testing the classification models.

Use Cases

The user will install the model_evaluator python package and import it into a project. The user will import the evaluate_regression or evaluate_classification methods for the classification and regression evaluation and provide the file name and optional parameters to the functions.

The software will authomatically run through all the regression or classification models.

The evaluate_classification function will return a data frame with the list of all classification algorithms tested against the dataset with their respective accuracy score. The algorithm with the maximum accuracy score is best suited for the supplied dataset. Similarly, the evaluate_regression function returns a data frame with the list of all regression models with some metrics. The model with the maximum R squared wins the day the provided data.

The package works for all datasets regardless of the number of features. Your features must be on the first columns. The dependent variable should always be the last column. The current package does not support categorical variables; however, there is a function to handle missing observations.