Implementing Data Cleaning and Transformation



Janani Ravi CO-FOUNDER, LOONYCORN www.loonycorn.com

Overview

Reading and exploring a dataset

Building a simple regression model

Performing univariate imputation of missing values

Performing multivariate imputation of missing values

Performing feature imputation and classification using a pipeline

Missing Data Deletion Imputation

Imputation

Fill in missing column values, rather than deleting records with missing values. Missing values are inferred from known data.

Imputation



Methods range from very simple to very complex

Simplest method: Use column average

Can interpolate from nearby values

Can even build model to predict missing values

Loading, cleaning, and exploring a dataset

Regression analysis using the automobile MPG dataset

Univariate feature imputation using the SimpleImputer in scikit-learn

Multivariate feature imputation using the IterativeImputer in scikit-learn

Indicate missing values using scikitlearn's missing value indicator

Perform feature imputation and classification using a scikit-learn pipeline

Summary

Reading and exploring a dataset

Building a simple regression model

Performing univariate imputation of missing values

Performing multivariate imputation of missing values

Performing feature imputation and classification using a pipeline