Missing Imputation with Scikit-learn: SimpleImputer()

Scikit-learn: supported imputations

- Mean / median imputation
- Arbitrary number imputation
- Frequent category imputation
- Imputation with bespoke string, i.e., “missing”
- Missing indicators

Scikit-learn: grid search

Grid search to identify most suitable imputation method.

Scikit-learn: subset of features

- ColumnTransformer

SimpleImputer(): Advantages

- Simple to use if applied to the entire dataframe
- Maintained by the Scikit-learn developers: good quality code
- Fast computation (it uses NumPy for calculations)
- Allows for grid search over the various imputation techniques
- Allows for different missing values encodings (you can indicate if the missing values are np.nan, or zeroes, etc.)

SimpleImputer(): Limitations

- Returns a NumPy array instead of a pandas dataframe, inconvenient for data analysis
- Needs to use additional classes to select which features to impute:
    - requires more lines of code
    - not so straightforward to use anymore.

Train In Data

THANK YOU

www.trainindata.com