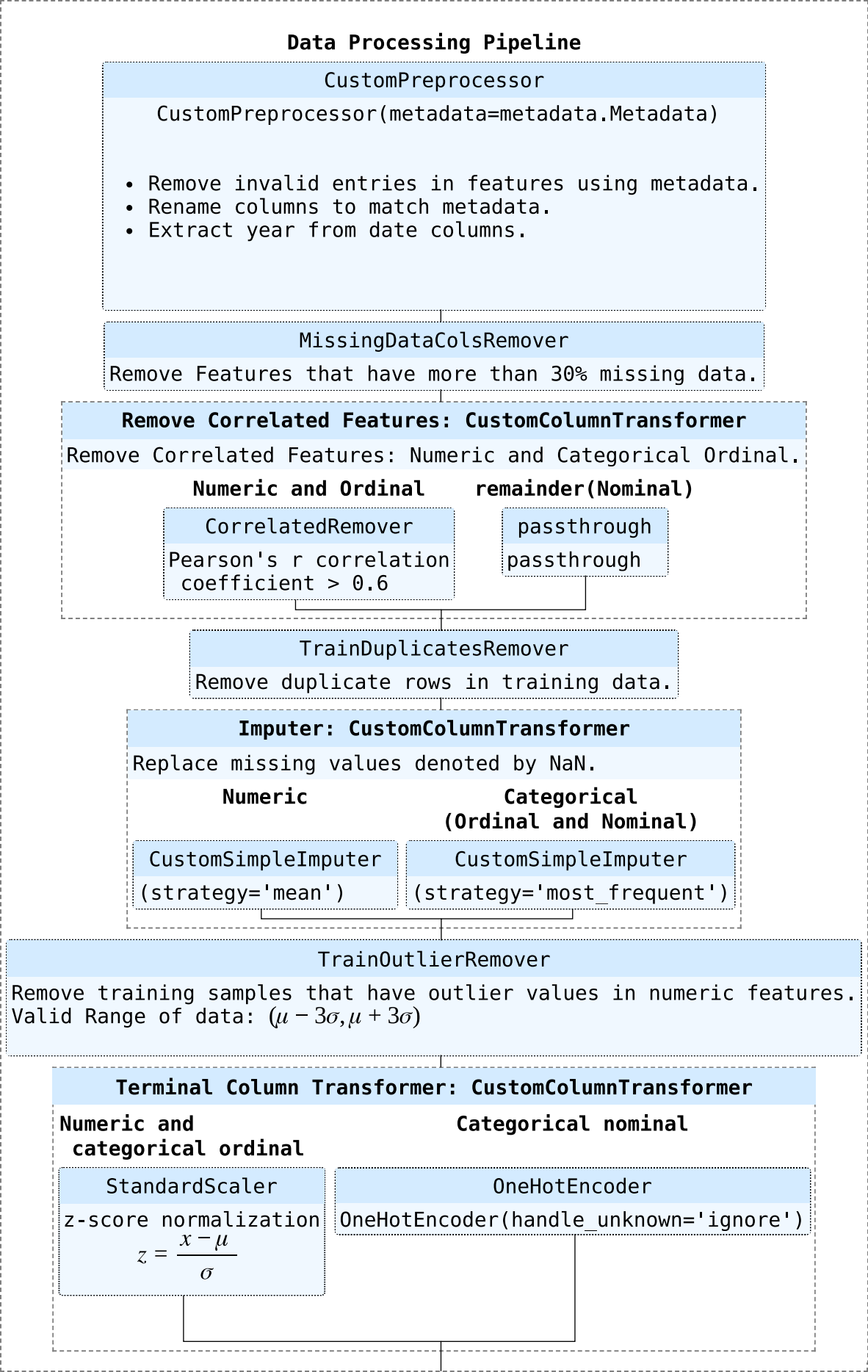
Data Processing Pipeline
CustomPreprocessor
CustomPreprocessor(metadata=metadata.Metadata)
• Remove invalid entries in features using metadata.
• Rename columns to match metadata.
• Extract year from date columns.
MissingDataColsRemover
Remove Features that have more than 30% missing data.
Remove Correlated Features: CustomColumnTransformer
Remove Correlated Features: Numeric and Categorical Ordinal.
Numeric and Ordinal
remainder(Nominal)
CorrelatedRemover
Pearson's r correlation coefficient > 0.6
passthrough
passthrough
TrainDuplicatesRemover
Remove duplicate rows in training data.
Imputer: CustomColumnTransformer
Replace missing values denoted by NaN.
Numeric
Categorical (Ordinal and Nominal)
CustomSimpleImputer
(strategy='mean')
CustomSimpleImputer
(strategy='most_frequent')
TrainOutlierRemover
Remove training samples that have outlier values in numeric features.
Valid Range of data: $(\mu - 3\sigma, \mu + 3\sigma)$
Terminal Column Transformer: CustomColumnTransformer
Numeric and categorical ordinal
Categorical nominal
StandardScaler
z-score normalization
$z = \frac{x - \mu}{\sigma}$
OneHotEncoder
OneHotEncoder(handle_unknown='ignore')