Feature Engineering: Takeaways

by Dataquest Labs, Inc. - All rights reserved © 2023

Syntax

- Create a `SimpleImputer` object to use in univariate imputation:

```
from sklearn.impute import SimpleImputer
imp = SimpleImputer(missing_values = np.nan,
                    strategy = "mean")
```

- Create a `KNNmputer` object to use in K-Nearest Neighbor imputation:

```
from sklearn.impute import KNNImputer
imp = KNNImputer(missing_values = np.nan,
                 n_neighbors=3)
imputed_X = imp.fit_transform(X)
```

- Calculate the quartiles of a set of values (box plot method):

```
percentiles = [0.25, 0.5, 0.75]
data_quartiles = np.percentile(data, percentiles)
```

- Calculate the Z-score for a set of values:

```
mhv_mean = housing["median_house_value"].mean()
mhv_std = housing["median_house_value"].std()
zscores = (housing["median_house_value"] - mhv_mean) / mhv_std
```

Concepts

- **Feature Engineering** is the process of extracting features from the data and transforming it into a format that the model can better understand or use.
- **Imputation** is the process of substituting missing data with other values, typically following some sort of strategy to choose these values.
- **Outlier Detection** refers to the process of detecting any **outliers** (data points that lie far from the rest of the obervations) and deciding how to handle them before the training process.
- **Box Plots** are visualizations based on the quartiles of a set of values, and can be used to identify outliers.
- **Z-scores** measure how far a data point is from the "average" data point in terms of standard deviations. It's based on the Normal or Gaussian distribution, which are known to contain 99% of their data within three standard deviations from the mean.
- **Downsampling** is the process of randomly selecting samples from the majority class and deleting them from the training dataset so that the minority class takes up a greater proportion of the data.
- **Upweighting** consists in making "copies" of the minority class to create more datapoints to balance the dataset.

Resources

- `scikit-learn` official documentation

- `SimpleImputer` class
- `KNNImputer` class
- `scikit-learn` vignette on imputation
- `boxplot()` method
- `percentile()` function
- Z-score
- `LogisticRegression` class

Takeaways by Dataquest Labs, Inc. - All rights reserved © 2023