Managing missing features

Sometimes a dataset can contain missing features, so there are a few options that can be taken into account:

- Removing the whole line
- Creating sub-model to predict those features
- Using an automatic strategy to input them according to the other known values

The first option is the most drastic one and should be considered only when the dataset is quite large, the number of missing features is high, and any prediction could be risky. The second option is much more difficult because it's necessary to determine a supervised strategy to train a model for each feature and, finally, to predict their value. Considering all pros and cons, the third option is likely to be the best choice. scikit-learn offers the class `Imputer`, which is responsible for filling the holes using a strategy based on the mean (default choice), median, or frequency (the most frequent entry will be used for all the missing ones).

The following snippet shows an example using the three approaches (the default...

Previous Section (/book/big_data_and_business_intelligence/9781785889622/3/ch03lvl1s

Next Section (/book/big_data_and_business_intelligence/9781785889622/3/ch03lvl1sec2