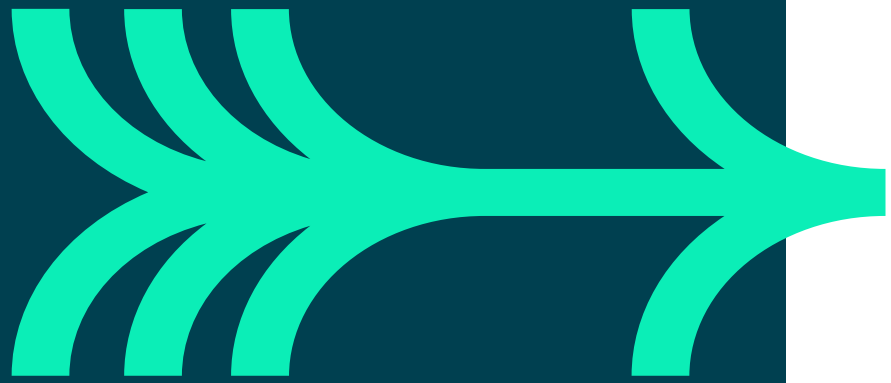
WHAT IS DATA PREPARATION?

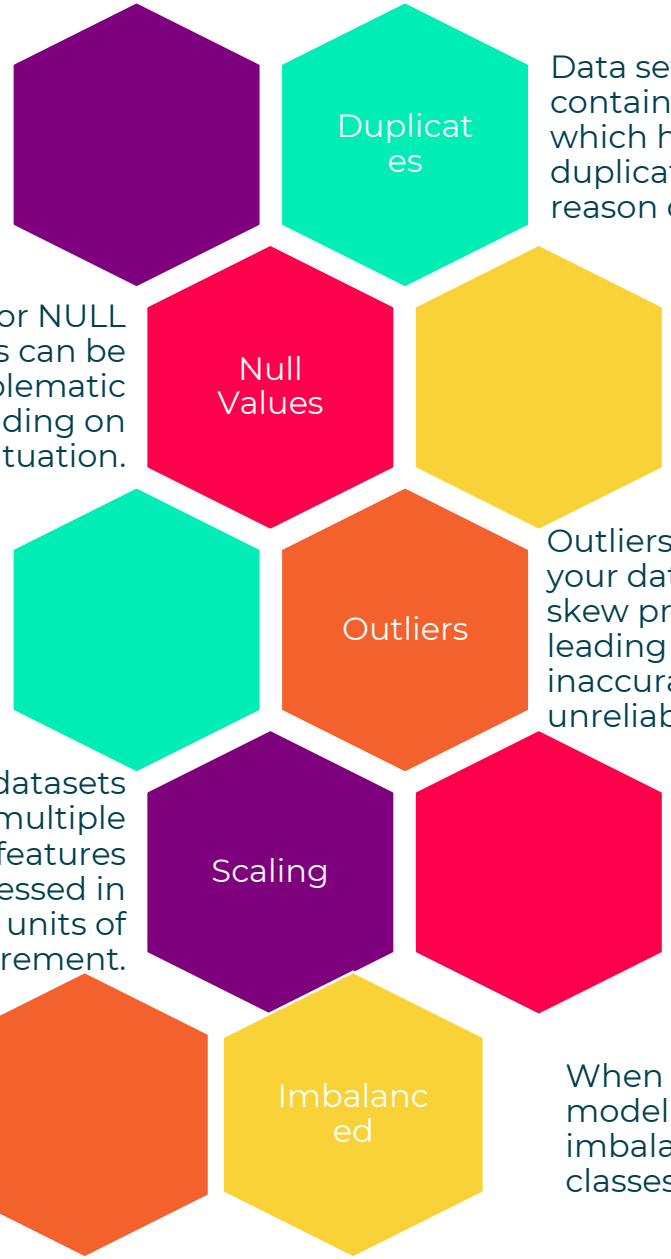
It is one of the most time-consuming and crucial processes in data mining. In simple words, data preparation is the method of collecting, cleaning, processing and consolidating the data for use in analysis. It enriches the data, transforms it and improves the accuracy of the outcome



DATA UNDERSTANDING

- Determining the distribution (Discrete/Categorical or Continuous)
- Population of values or identification of missing values (dense or sparse)
- Generating a statistical profile of the data (Min, Max, Mean, Counts, Distinct Count, etc)
- Identifying correlation within the data set

PRE-PROCESSING



Data sets can contain records which have been duplicated for one reason or another.

Missing or NULL values can be problematic depending on the situation.

Outliers within your data set can skew predictions leading to an inaccurate or unreliable model.

Most datasets contain multiple numeric features expressed in different units of measurement.

When building a classification model we may have an imbalanced distribution of classes – Anomaly Detection

MISSING DATA

How do we deal with missing data?

There are three main options:

1) Removal

2) Imputation – Requires skill

3) Leave as is, some models can deal with missing values

TYPES OF MISSING DATA

Missing completely at random (MCAR)

- data are missing independently of both observed and unobserved data.

Missing at random (MAR)

- given the observed data, data are missing independently of unobserved data.

Missing Not at Random (MNAR)

- missing observations related to values of unobserved data.

MISSING VALUES

How are we going to handle them?

- Make them up?
- Do nothing?
- Remove the rows that contain them?
- Does this then bias the sample?
- Fix the source system?
- Replace the value with the mean, the mode?

OUTLIERS

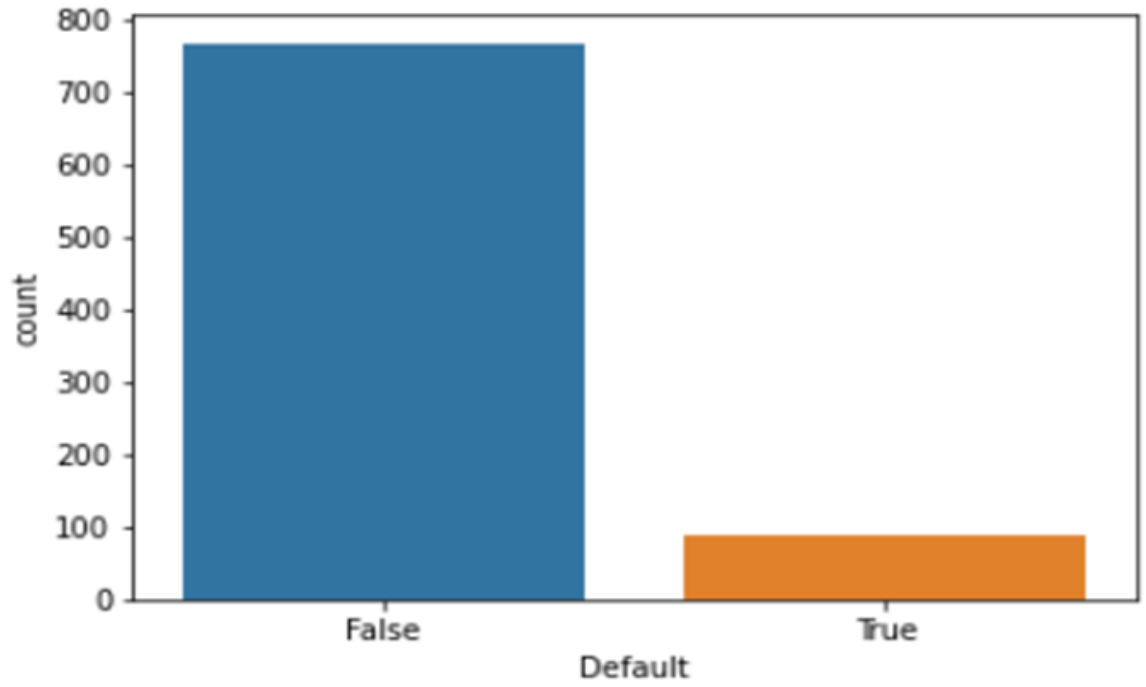
- How are we going to handle them?
- Do nothing?
- Verify them?
- Remove the rows that contain them?
- Does this then bias the sample?
- Are they symmetrically disposed around the mean?
- Fix the source system?
- Replace the value with the mean, the mode?
- Bin the values and have a category such as > 50?

IMBALANCED DATA FOR CLASSIFIERS

When the occurrence of one class outweighs another.

Always the case with Anomaly Detection

```
:  1  sns.countplot(x='Default', data = df);
```



DEMO AND EXERCISE

Demo and exercise with basic data prep dealing with Nans and outliers.

We will cover imbalanced classes in logistic regression.

The data prep done now will feed into the Linear Regression

Open the notebook in this folder for Trainer Demo and exercises