L174031 Samra Fakhar

L174162 Nuzha Khaled

Data Munging

Data Munging is the process of transforming and mapping data from one raw data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.

Raw data consists of missing values and outliers. In datasets, missing values could be represented as '?', 'nan', 'N/A', blank cell, or sometimes '-999', 'inf', '-inf'

- MCAR means that the occurrence of missing values is completely at random, not related to any variable.
- MAR implies that the missingness only relate to the observed data
- NMAR refers to the case that the missing values are related to both observed and unobserved variable and the missing mechanism cannot be ignored.

These need to be handled before the data can be used for useful purposes. We should estimate those values wisely depending on the amount of missing values and the expected importance of variables.T here are many ways to do it.

Imputation

Imputation simply means replacing the missing values with some guessed/estimated ones.

These include

1. Mean, median, mode imputation

A simple guess of a missing value is the mean, median, or mode (most frequently appeared value) of that variable.

2. Regression imputation

If we know there is a correlation between the missing value and other variables, we can often get better guesses by regressing the missing variable on other variables

3. K-nearest neighbor imputation

K-nearest neighbour (KNN) imputation is an example of neighbour-based imputation. For a discrete variable, KNN imputer uses the most frequent value among the k nearest neighbours and, for a continuous variable, use the mean or mode.

4. Last observation carried forward

using last valid observation to fill the NA's is known as Last observation carried forward (LOCF)

5. Interpolation

if we are dealing with time-series data, it might make senes to use interpolation of observed values before and after a timestamp for missing values

6. Predictive mean matching

It combines the idea of model-based imputation (regression imputation) and neighbor-based (KNN imputer). First, the predicted value of target variable Y is calculated according to a specified model and a small set of candidate donors are chosen from complete cases that have Y close to the predicted value. Then, a random draw is made among the candidates and the observed Y value of the chosen donor is used to replace the missing value.

Removing Missing Values

Removing values from the data where any cell value is missing is not advisable. This is because this may result in losing other important information from the data. It may work only in the case if the missing values are very less as compared to the total data

Which method to use?

when deciding how to impute missing values in practice, it is important to consider:

- the context of the data
- amount of missing data
- missing data mechanism

Data Exploration

It can be done by

- looking at few top rows by using the function head()
- looking at more rows by printing the dataset
- looking at summary of numerical □elds by using describe() function. describe() function would provide count, mean, standard deviation (std), min, quartiles and max in its output. Using this we can analyze the missing values in our data.
- For the non-numerical values, we can look at frequency distribution to understand whether they make sense or not.

Distribution analysis

By plotting histograms and box plots and bar charts, we can understand the distributions. This conirms the presence of of outliers/extreme values.