Data Cleaning & Preparation

Intro

Imagine the initial stages of a machine learning project for preparing a meal for your family. The beginning stages of cooking your meal involves cleaning the kitchen, gathering your ingredients, washing vegetables, and then you start cooking your meal, which can be a messy process itself. Analogously data cleaning is gathering the ingredients and washing the vegetable stage. Before you gather insights, perform feature engineering, and train models on the data, you know the fun stuff, you must clean and massage the data first.

Terminology

1. **Imputation**: In machine learning, imputation is the formal word for replacing missing values.
2. **NaN / NA:** In datasets, *NaN* or *NA* is used to represent missing values. *NaN* means *Not A Number*, and *NA* stands for *N*ot *A*vailable.

---

1. Imputation

Data for machine learning related tasks is often messy, with lots of missing values. Imputation is the process of removing the missing data, filling in the missing data, or in rare cases dropping the *feature* all-together. Most of the time, a **NaN or NA** signifies a missing value, but **negative/positive infinity** is also standard. Other times, missing data is marked by values that do not make sense for that feature. For example, a person's age is negative to signify missing data.

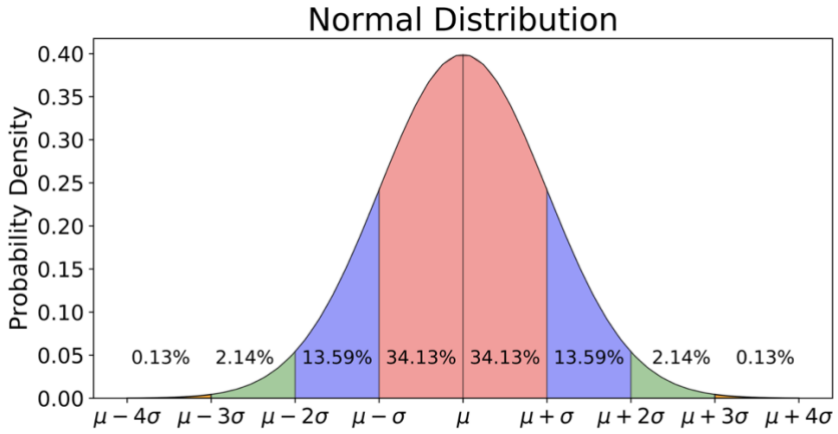
1. **Removing the Missing Data**

If the data that is missing is entirely random, then eliminating the rows that have missing information is a quick way to fix this issue. Note that removing rows that have missing data reduces the training size. If you have lots of missing data, then this is not a sound approach; however, if the dataset is quite large, then removing missing data is perfectly ok.

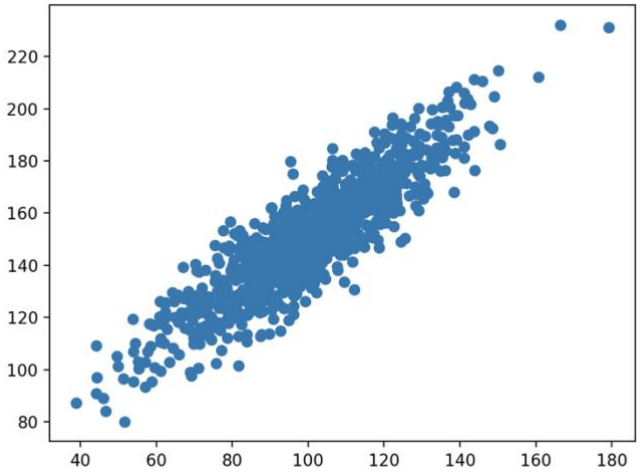
2. **Filling in the Missing data**

Sometimes missing data gives clues as to what those values are. For example, if you have a boolean feature full of 1's and NaN's, then presumably, the missing values are 0's. In other cases, plotting the distribution of the data provides insights on how to best fill the missing values. Uniformly distributed data are all the same value. For categorical data, if the values are uniformly distributed, then filling the values with 'Other' is done. This keeps the uniform

distribution of the categorical data. In the case of normally distributed data, we know that 68.2% of the data falls within one standard deviation of the mean. If the data is normally distributed, we use this property to replace missing values with the mean.



Another strategy is to analyze correlation amongst features. For example, if feature *A* and feature *B* have a linear relationship, then the missing values in *B* can be extrapolated from *A.*



If all else fails, a fail-safe method is to fill the missing values with the median of the feature. The median, unlike the mean, is agnostic to outliers.

In some cases, the data provided can be incredibly sparse, with a large amount of data missing. From my personal experience, I have seen data that have a column with over 90% of the data missing. In these cases, dropping the feature altogether is the solution.