Missing Values Mechanism and Imputation

Data may be missing:

1. Completely at random;
2. Missing randomly, but not completely randomly;
3. Missing due to some unmeasured (confounding) reason;
4. Missing due to the missing value itself.

Depending on the reason for the missing data, it may be a reasonable choice to use the other variables to predict the missing data. However, there are situations where doing this is going to exacerbate problems caused by the missing data. Imputing data with measures of central tendency generally isn't recommended, and methods such as Multiple Imputation by Chained Equations will be better, although they tend to take a while to run. But with as many methods for imputing missing data available as there are, there really is no reason to simply replace missing data with the mean of the vector. As one of my stats professors once told me, "When data are missing, they're missing." (sort of a Yogi Berra-ism, but it has a certain wisdom about it because it is impossible to that data are missing at random).

 I used RMSE as the metric and tried various Models (KNN, SVM etc.) to impute the values, improved my local CV a bit, but not anything significant. But there are many features where imputing -1 makes far more sense, i.e. the feature does not exist,

At the same time, not all the models can handle -1 values; XGBoost can!

The basic assumption of using MICE (Multivariate imputation by chained equations)

method is that the data was missing at random, otherwise it will result in biased estimates. MICE operates under the assumption that given the variables used in the imputation procedure, the missing data are Missing At Random (MAR), which means that the probability that a value is missing depends only on observed values and not on unobserved values.

How does MICE work - example?

To make the chained equation approach more concrete, imagine a simple example where we have 3 variables in our dataset: age, income, and gender, and all 3 have at least some missing values. The MAR assumption would imply that the probability of a particular variable being missing depends only on the observed values, and that, for example, whether someone’s income is missing does not depend on their (unobserved) income. In step 1 of the MICE process, each variable would first be imputed using, e.g., mean imputation, temporarily setting any missing value equal to the mean observed value for that variable. Then in step 2 the imputed mean values of age would be set back to missing. In step 3, a linear regression of age predicted by income and gender would be run using all cases where age was observed. In step 4, predictions of the missing age values would be obtained from that regression equation and imputed. At this point, age does not have any missingness. Steps 2–4 would then be repeated for the income variable. The originally missing values of income would be set back to missing and a linear regression of income predicted by age and gender would be run using all cases with income observed; imputations (predictions) would be obtained from that regression equation for the missing income values. Then, steps 2–4 would again be repeated for the variable gender. The originally missing values of gender would be set back to missing and a logistic regression of gender on age and income would be run using all cases with gender observed; predictions from that logistic regression model would be used to impute the missing gender values. This entire process of iterating through the 3 variables would be repeated until convergence; the observed data and the final set of imputed values would then constitute one “complete” data set.

Let's think about an example where you want to estimate the mean income within a certain population, which you obtain via questionnaires. Suppose some income measures are missing. When missingness is MCAR, then the missingness is completely random, as if some questionnaires were lost by chance. When missingness is MAR, than the missingness is random within subgroups of other observed variables, for instance, suppose that you also collect data on the profession of subject and that say mangers are more likely not the share their income, than, within subgroups of the profession, missing is random. Missingness is MNAR when n the reason for missingness depends on the missing values themselves/ For instance, suppose people don't want to share their income when it is below 1000 euros per month because they are ashamed of it. I hope this enhances your understanding concerning these definitions.

The “modern” missing data procedures are

(a) the EM algorithm

(b) multiple imputation under the normal model

(c) ML methods, often referred to as full-information maximum likelihood (FIML) methods.