Study of the effectiveness of selected data imputation techniques on the mobile phone parameters dataset

Dawid Chomicz

2022-06-29

The aim of the study is to compare the effectiveness of selected data imputation techniques. For this purpose, the *phone.csv* data set describing the technical parameters of mobile phones was used. The process of generating data deficiencies and imputating these deficiencies was simulated, and then the average descriptive statistics for each imputation technique used were estimated. The calculations were made with the R programming language.

Characteristics of the dataset

The data comes from the resources of the kaggle.com site. The set contains 1000 observations in the form of numerical variables. Detailed description of each variable:

- *battery* — total battery capacity in mAh,
- *speed* – processor speed,
- *fc* – front camera megapixels,
- *memory* – internal memory in GB,
- *weight* – phone weight,
- *pc* – main camera megapixels,
- *ram* – RAM in MB,
- *sc_h* – screen height in cm,
- *sc_w* – screen width in cm,
- *talk_t* – maximum usage time in hours without recharging the battery.

Details of the simulation process

The performed simulation aims to compare the effectiveness of selected data imputation techniques on the same set of observations. For this purpose, an algorithm consisting of three steps was prepared:

1. Generating MCAR (*missing completely at random*) type missing data in the set, covering 20% of the data without taking into account the first two variables.
2. Application of the following imputation techniques to the generated set: mean, median, regression, K nearest neighbors, random forest.

3. Calculation of statistics describing the structure of the dataset after imputation for each technique.

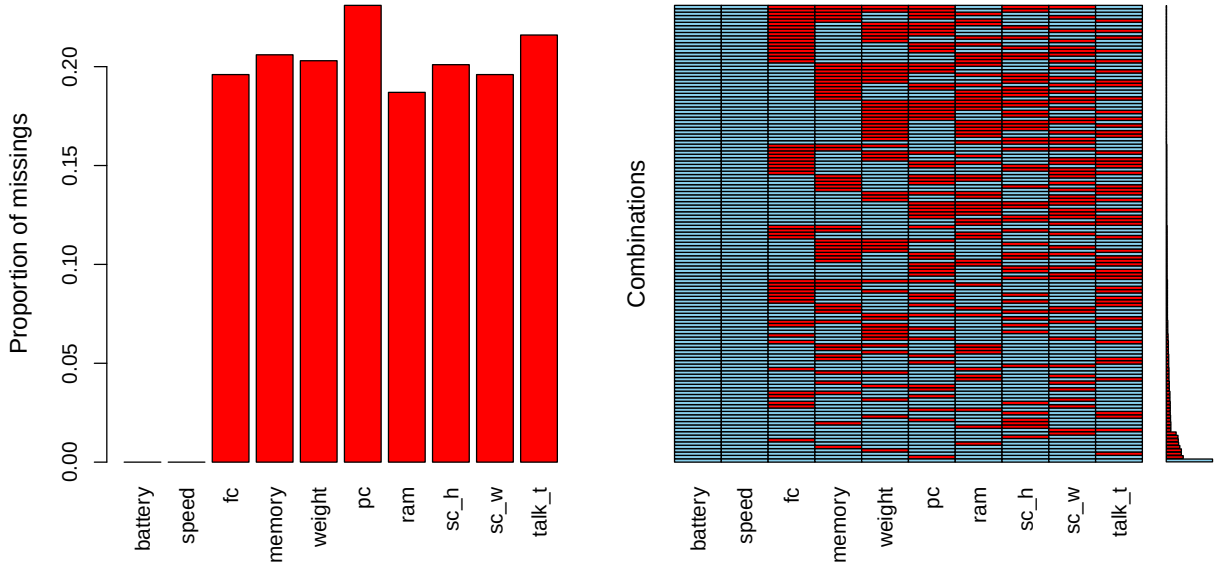
The above algorithm was iterated a thousand times and the results of these iterations were averaged.

Parameters adopted for individual techniques:

- Regression – generalized linear model with automatic selection of the linking function and distribution of the modeled variable, where the dependent variables were all variables with missing values, and the explanatory variables were the first two variables,
- K najbliższych sąsiadów – parameter $k = 5$,
- Las losowy – the formula for modeled and explanatory variables is the same as in regression.

Missing data characteristics

Figure 1. The frequency and distribution of missing data



In the chart above, you can see the frequency and distribution of missing data in the generated dataset. The frequency as assumed fluctuates around 20%. There are no deficiencies in the first two variables. Combinations od missing data do not show strong relations, although there is a slightly greater number of deficiency combinations between *fc* and *memory* and between *memory* and *weight* than in other cases.

Simulation results

The first table below lists the descriptive statistics for the initial, unchanged dataset. The following tables show the subtraction of the averaged statistics for simulated imputation techniques from the base set statistics. The * abs_sum * column contains the sum of the absolute difference values for each statistic.

For initial data

	battery	speed	fc	memory	weight	pc	ram	sc_h	sc_w	talk_t
x_sr	1248.51	1.54	4.59	33.65	139.51	10.05	2139.00	11.99	5.32	11.09
Me	1246.50	1.50	3.00	34.50	139.00	10.00	2153.50	12.00	5.00	11.00
s	432.46	0.83	4.46	18.13	34.85	6.10	1088.09	4.32	4.24	5.50
Vs	0.35	0.54	0.97	0.54	0.25	0.61	0.51	0.36	0.80	0.50
s_x_sr	13.68	0.03	0.14	0.57	1.10	0.19	34.41	0.14	0.13	0.17
g1	0.04	0.19	0.99	-0.07	0.01	0.00	-0.05	-0.04	0.78	0.02

Mean imputation

	fc	memory	weight	pc	ram	sc_h	sc_w	talk_t	abs_sum
x_sr	0.00	-0.01	0.02	0.00	0.18	0.00	0.01	0.00	0.22
Me	-1.59	0.84	-0.49	-0.06	14.69	0.00	-0.31	-0.09	18.06
s	0.47	1.93	3.69	0.64	114.91	0.46	0.45	0.58	123.13
Vs	0.10	0.06	0.03	0.06	0.05	0.04	0.08	0.05	0.48
s_x_sr	0.01	0.06	0.12	0.02	3.63	0.01	0.01	0.02	3.89
g1	-0.12	0.01	0.00	0.00	0.01	0.00	-0.09	0.00	0.23

Median imputation

	fc	memory	weight	pc	ram	sc_h	sc_w	talk_t	abs_sum
x_sr	0.23	-0.16	0.11	0.01	-2.41	0.00	0.14	0.02	3.09
Me	-0.43	0.05	-0.06	0.01	1.70	0.00	0.38	0.01	2.65
s	0.44	1.92	3.68	0.64	114.87	0.46	0.44	0.58	123.03
Vs	0.05	0.06	0.03	0.06	0.05	0.04	0.06	0.05	0.40
s_x_sr	0.01	0.06	0.12	0.02	3.63	0.01	0.01	0.02	3.89
g1	-0.26	0.04	-0.01	-0.01	0.01	0.01	-0.19	-0.01	0.53

Regression imputation

	fc	memory	weight	pc	ram	sc_h	sc_w	talk_t	abs_sum
x_sr	0.00	-0.01	0.02	0.00	0.18	0.00	0.01	0.00	0.22
Me	-1.48	0.75	-0.44	-0.02	12.44	-0.01	-0.12	-0.03	15.28
s	0.47	1.92	3.67	0.64	114.73	0.45	0.45	0.57	122.92
Vs	0.10	0.06	0.03	0.06	0.05	0.04	0.08	0.05	0.48
s_x_sr	0.01	0.06	0.12	0.02	3.63	0.01	0.01	0.02	3.89
g1	-0.12	0.01	0.00	0.00	0.01	0.00	-0.09	0.00	0.23

K nearest neighbors imputation

	fc	memory	weight	pc	ram	sc_h	sc_w	talk_t	abs_sum
x_sr	0.11	-0.02	-0.11	0.10	-3.20	0.04	0.05	-0.01	3.65
Me	-0.67	0.14	-0.32	0.08	-2.97	0.00	0.27	0.02	4.47
s	0.27	1.12	2.04	0.32	61.13	0.22	0.24	0.31	65.66
Vs	0.04	0.03	0.01	0.03	0.03	0.02	0.04	0.03	0.22
s_x_sr	0.01	0.04	0.06	0.01	1.93	0.01	0.01	0.01	2.08
g1	-0.05	0.01	0.00	-0.03	0.01	-0.01	-0.03	0.00	0.15

Random forest imputation

	fc	memory	weight	pc	ram	sc_h	sc_w	talk_t	abs_sum
x_sr	-0.01	-0.01	-0.01	0.00	0.57	0.00	0.00	0.00	0.59
Me	-1.00	0.32	-0.58	0.02	4.94	0.00	0.01	-0.01	6.89
s	0.42	1.72	3.30	0.58	102.12	0.40	0.40	0.51	109.45
Vs	0.09	0.05	0.02	0.06	0.05	0.03	0.07	0.05	0.43
s_x_sr	0.01	0.05	0.10	0.02	3.23	0.01	0.01	0.02	3.46
g1	-0.08	0.01	0.00	0.00	0.00	0.00	-0.06	0.00	0.15

The table below compares all the sums of the absolute values of the estimated differences for each simulated imputation technique.

	mean	median	regression	kNN	random forest
x_sr	0.22	3.09	0.22	3.65	0.59
Me	18.06	2.65	15.28	4.47	6.89
s	123.13	123.03	122.92	65.66	109.45
Vs	0.48	0.40	0.48	0.22	0.43
s_x_sr	3.89	3.89	3.89	2.08	3.46
g1	0.23	0.53	0.23	0.15	0.15

Comparing the above tables, it can be concluded that:

In the case of the arithmetic mean for the studied variables, imputations with regression and random forest turned out to be the closest.

In the case of the median for the studied variables, the imputations with the median and the K nearest neighbors turned out to be the most similar.

In the case of the standard deviation for the studied variables, the imputations of K nearest neighbors and the random forest turned out to be the most similar.

In the case of the coefficient of variation for the studied variables, the imputations of K nearest neighbors and the random forest turned out to be the most similar.

In the case of the standard error of the mean for the studied variables, the imputations of K nearest neighbors and the random forest turned out to be the closest.

In the case of the asymmetry coefficient for the studied variables, the imputations of K nearest neighbors and random forest turned out to be the closest.

Conclusion: The imputation technique that produces results on average closest to the original set values is the K nearest neighbors. In the case of the mean and the median, better results were achieved by the mean and median imputations respectively, which seems natural due to the specificity of these measures.