Dropping Incomplete Records is (not so) Straightforward

Rianne M. Schouten, Victoria Taşcău(✉), Gabriel G. Ziegler, Davide Casano, Marco Ardizzone, and Michael-Angelos Erotokritou

Eindhoven University of Technology, Eindhoven, The Netherlands
r.m.schouten@tue.nl, {v.tascau,g.gomes.ziegler,d.casano,m.ardizzone, m.a.erotokritou}@student.tue.nl

Abstract. A straightforward approach to handling missing values is dropping incomplete records from the dataset. However, for many forms of missingness, this method is known to affect the center and spread of the data distribution. In this paper, we perform an extensive empirical evaluation of the effect of the drop method on the data distribution. In particular, we analyze two scenarios that are likely to occur in practice but are not often considered in simulation studies: 1) when features are skewed rather than symmetrically distributed and 2) when multiple forms of missingness occur simultaneously in one feature. Furthermore, we investigate implications of the drop method for classification accuracy and demonstrate that dropping incomplete records is doubtful, even when test cases are dropped as well.

Keywords: Missing data · dropping incomplete records · skewness

1 Introduction

A straightforward approach to handling missing values is dropping incomplete records from the dataset [3]. For some forms of missingness, this method is known to affect the data distribution by creating a shift in the mean of the distribution and by influencing its standard deviation. In other situations, dropping incomplete records merely reduces the dataset size, although this could bring about new problems such as imprecise statistical estimates or lack of training data [11,16].

For the student, scientist or engineer, dropping incomplete records allows to quickly move forward with developing the desired machine learning model. However, such a model may not live up to its expectations, albeit because after deployment incomplete cases that are dropped cannot be predicted nor classified. At the same time, incomplete training data could make the development of an AI system conceptually or practically impossible [14].

In this paper, we perform an extensive empirical evaluation of the effects of missing data. The goal of our investigation is twofold. First, we add to existing knowledge by studying how measures of the center and spread of the data

B. Crémilleux et al. (Eds.): IDA 2023, LNCS 13876, pp. 379–391, 2023.
https://doi.org/10.1007/978-3-031-30047-9_30

distribution are affected when 1) features are skewed rather than symmetrically distributed and 2) several forms of missingness occur simultaneously rather than separately. Both situations are likely to occur in practice but are often not considered in analyses of missing data problems.

Second, we investigate implications of the drop method for classification accuracy. In the study of missing data, incomplete datasets are generally randomly split into training and test data [7–9]. Although such an approach does justice to a development process that should align with the situation after deployment, it prohibits the direct investigation of effects of missing data and only allows the study of imputation methods. On the other hand, when test data is a specific selection of complete records in an incomplete dataset, the distribution of the training data may differ from test or application data, creating issues such as concept drift [22]. In this paper, we demonstrate how missing data shifts the observed data distribution, that some forms of missingness behave unexpectedly when the distribution is skewed and that classification accuracy is affected whether or not you drop incomplete test records.

2 Background

2.1 Preliminaries

We consider a d-dimensional space $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_{d-1}, \mathcal{Y}\}$ and let $\mathbf{X} = (X_1, X_2, ..., X_{d-1}, Y)$ be a random variable taking values in $\mathcal{X}$. Note that we write X_j and Y to distinguish predictor variables from the assigned outcome variable, but we assume that all variables have a joint distribution $P(\mathbf{X})$. Next, we define a *complete* dataset $D \in \mathbb{R}^{n \times d} = \{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^n\}$ to be a collection of n independent and identically distributed realizations of $\mathbf{X}$.

Furthermore, we define a missing data indicator $R \in \{0, 1\}^{n \times d}$ that reveals whether values in D are missing or not. Here, $r_{ij} = 1$ when d_{ij} is observed and $r_{ij} = 0$ when d_{ij} is missing for all $i \in \{1, 2, ..., n\}$ and $j \in \{1, 2, ..., d\}$. We distinguish the hypothetically complete dataset D from the masked, or *incomplete*, dataset by denoting the latter with by $\tilde{D} = \{\tilde{\mathbf{x}}^1, \tilde{\mathbf{x}}^2, ..., \tilde{\mathbf{x}}^n\}$.

In this paper, we investigate the effect of dropping incomplete records for various missing data scenarios. Essentially, the procedure discards all observations that have at least one missing value. We denote the resulting dataset by $\bar{D} = \{\tilde{\mathbf{x}}^i | \mathbf{r}^i = \mathbf{1}\}$; the vector of indications for case i should be an all-ones vector. Denoting the sample size of the dropped dataset by $\bar{n}$, the overall missingness percentage is defined as $\rho = 100\frac{\bar{n}}{n}$.

2.2 Missing Data Mechanisms

In the study of missing data, the process that governs the probability that certain values are missing is called the *missing data model* or *missing data mechanism* [3,11,15]. It is helpful to understand which missing data mechanisms are present in order to choose appropriate missing value treatments. Rubin [15] distinguishes the following three missing data mechanisms.

First, data is said to be Missing Completely At Random (MCAR) if the probability of being missing is unrelated to observed and missing data distribution: $P(R|D_{\text{obs}}, D_{\text{mis}}, \psi) = P(R|\psi)$. With MCAR every data value has the same, fixed probability of being missing, denoted by ψ. Consequently, observed and missing data distributions will be similar, and a method such as dropping incomplete records will allow for the estimation of unbiased statistical parameters [3].

Second, data is Missing At Random (MAR) if observed data governs the missingness probabilities: $P(R|D_{\text{obs}}, D_{\text{mis}}, \psi) = P(R|D_{\text{obs}}, \psi)$. Here, observed and missing data distribution may be different and statistical inferences based on observed data alone may be severely biased. However, under the MAR assumption, observed data contains all information necessary to model the missing data; $P(D_{\text{mis}}|D_{\text{obs}}, R)$. This concept of *ignorability* is an important starting point for many imputation methods [5].

When data is neither MCAR nor MAR, information about the missing values is missing from the dataset. Then, data is Missing Not At Random (MNAR). In other words, the probability to be missing depends on the missing value itself: $P(R|D_{\text{obs}}, D_{\text{mis}}, \psi)$.

2.3 Missing Data Types

The missing data model can be any function that maps a numerical value to a probability: $f : x \mapsto p$, with $p \in [0, 1]$. In practice, when performing experiments with missing data, the logistic or sigmoid function $f_{\text{logistic}}(x) = \frac{1}{1+\exp^{-x}}$ is a convenient choice, partially because for any normally distributed input vector $\mathbf{x} = \{x^1, x^2, ..., x^n\}$, the sum of n Bernoulli trials with success probabilities $\mathbf{p} = \{p^1, p^2, ..., p^n\}$ equals $n\bar{p}$ with $\bar{p} = \frac{1}{n}\sum_{i=1}^{n} p^i = 0.5$ [13]. In practice, this means that 50% of the records will be incomplete.

Recently, [17] proposed a multivariate amputation procedure that allows for easy control of missing data characteristics such as missing data mechanisms, percentage and patterns (every unique row in R is considered a pattern). In addition, they distinguish four versions of the logistic function that allows the researcher to control what part of the data distribution will be masked. These versions are called missing data *types* and can be seen in Fig. 1.

The *right* missingness type is the normal logistic function and assigns high probabilities to large values: $f_{\text{right}}(x) = f_{\text{logistic}}(x)$. The opposite is the *left* missingness type: $f_{\text{left}}(x) = f_{\text{logistic}}(-x)$. Furthermore, *tail* and *mid* missingness assign high probabilities to values in the tails and center of the distribution respectively: $f_{\text{tail}}(x) = f_{\text{logistic}}(|x| - 0.75)$ and $f_{\text{mid}}(x) = f_{\text{logistic}}(-|x| + 0.75)$. Here, 0.75 is a fixed value that ensures $\rho = 50\%$ missingness (other percentages are easily obtained by shifting the logistic functions horizontally). All these missing data types reflect real-world scenarios such as survey questions not being answered for extreme values or medical tests not being executed for 'average' patients. In Sect. 5, we show that the effect of dropping incomplete records varies for these missingness types [17,18].

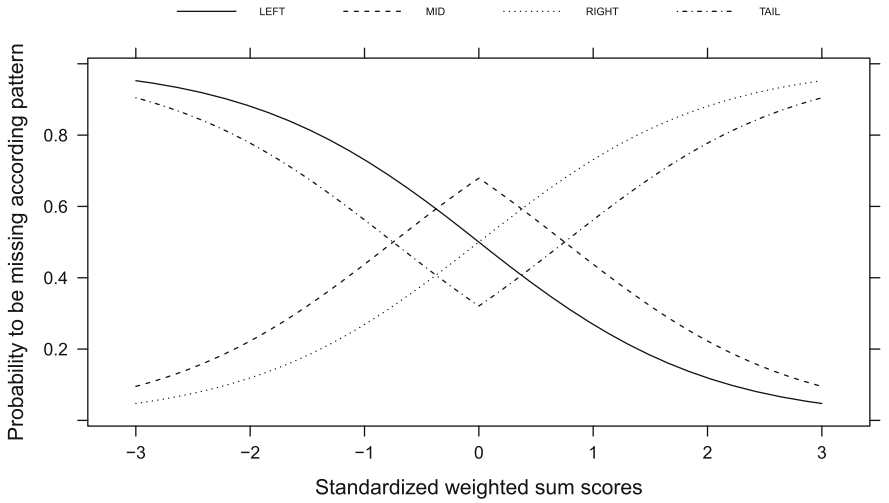

Fig. 1. Four missing data types according to [17]. Standardized weighted sum scores are linear combinations of observed data. In our notation, we use the general term $\mathbf{x}$.

3 Related Work

Handling missing values by dropping incomplete records is also known as complete-case analysis or listwise deletion. Especially in the domain of statistics, the effect of complete-case analysis on the validity of statistical estimates has been studied substantially [2,4,10,11,21]. In the machine learning domain, dropping incomplete records is generally accepted if missingness percentages are small or missing values are evenly divided over the data distribution [1,8]. However, it is not part of empirical studies simply because dropped test cases cannot be evaluated [7–9].

We consider our paper to build on work by [18]. Schouten et al. [18] investigate whether the correlation between data features influences the effect of missing data on estimates of the mean, standard deviation and correlation. They find that for an estimate of the mean, when data correlations are small, the effects of MAR missingness converge towards those of MCAR missingness; in contrast, for large data correlations MAR behaves like MNAR.

Furthermore, [18] compare the effects of right, left, tail and mid missingness types. They empirically show that right and left missingness affect the center of the distribution, whereas tail and mid missingness affect the spread of the distribution. The larger the data correlation, the more these effects appear. These results are interesting because they show that for some forms of MAR and MNAR missingness, depending on the statistical quantity of interest, dropping incomplete records may not be as harmful as we may think.

In this paper, we evaluate the behavior of missing data for two scenarios that are likely to occur in practice but have not been studied empirically: 1) the effect of skewness and 2) the simultaneous presence of multiple mechanisms. We furthermore investigate the drop method from a machine learning point of view by analyzing its effect on classification accuracy.

4 Experimental Design

We design two experiments. First, we perform a synthetic data experiment to investigate the effect of skewness. Thereafter, Sect. 4.1 outlines our experiments with a real-world, public dataset. All our experimental code and results are available at https://github.com/Research-Topics-in-Data-Mining/missingness-effect-complete-dataset.

Synthetic dataset generation is done by drawing a dataset H with $n = 10\,000$ observations from a multivariate normal distribution $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$ with mean vector $\mu_H = [10, 10]$ and covariance matrix $\Sigma_H = [1, 0.5; 0.5, 1]$. In a copy $H' = H$ we then create right-directed skewness in the first feature by squaring all values larger than 3 standard deviations from the center. Formally, for $i \in \{1, 2, ..., n\}$, $A^i = (H_1^i)^2$ if $H_1^i > f_{\text{mean}}(H_1) + 3f_{\text{std}}(H_1)$ and $A^i = H_1^i$ otherwise. Then, $H_1' = A^1 \cap A^2 \cap ... \cap A^n$. The amount of skewness can be calculated by $f_{\text{skew}}(D_j) = 3(f_{\text{mean}}(D_j) - f_{\text{median}}(D_j))/f_{\text{std}}(D_j)$ [13]. With our approach, the average skewness is 0.11.

To ensure fair comparison between datasets H and H', we standardize both datasets into D and D' respectively such that $\mu_D = \mu_{D'} = [0, 0]$ and $\Sigma_D = \Sigma_{D'} = I_2^{-1}$. Subsequently, for both D and D' separately, we generate $\rho = 50\%$ missingness in the first feature for all combinations of missing data mechanisms and the four missingness types; resulting in 9 scenarios: MCAR, 4 × MAR, 4 × MNAR. MAR missingness is created by using the observed values in the second feature. For the exact procedure, we apply the multivariate amputation procedure implemented in function `ampute` [17] in `R`.

We evaluate effects of missing data on the center and spread of the distribution by calculating the difference between the dropped and the complete dataset for two measures of the center, the mean and median, and two measures of the spread, the standard deviation and interquartile range. We do this for the skewed and non-skewed data separately. For instance, the mean shift for the non-skewed data is $\varphi_{\text{mean shift non-skewed}} = f_{\text{mean}}(D_1) - f_{\text{mean}}(\bar{D}_1)$. We repeat the experiment $T = 1\,000$ times.

4.1 Real-world Data Experiment

The real-world public Breast Cancer dataset[1] [12,20] contains 10 predictor features and 1 binary outcome variable for $n = 569$ cases. We generate a simple missing data pattern where missing values occur in one feature. Specifically, we

[1] https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic).

decide to ampute *smoothness* based on observed data in feature *symmetry*; the two features have a correlation coefficient of 0.6 and smoothness has a medium importance in a Random Forests (RF) classification model.

We investigate the effect of the simultaneous occurrence of missing data mechanisms. Note that such a concurrent existence can happen in several ways. For instance, multiple patterns exist where each pattern follows a different mechanism, or the missingness probabilities come from both observed and unobserved data. We choose an option where missing values occur in one feature; yet for each mechanism some other subsection of the data is used to determine which cases should be amputed. We use the implementation of the multivariate amputation procedure in Python: `pyampute` [19]).

Specifically, MCAR, MAR and MNAR mixtures are created by varying the missingness percentages for each mechanism $\rho_{\text{mcar}}, \rho_{\text{mar}}, \rho_{\text{mnar}} \in \{0, 10, 20\}$. Consequently, we obtain 26 configurations (the scenario 0-0-0 will not generate any missing values), where some configurations contain a single mechanism and others mixtures of 2 or 3 mechanisms. We sequentially perform the experiment for right, left, mid and tail missingness types (thus, a MAR-MNAR mixture follows the same type), and repeat every simulation scenario $T = 1\,000$ times. Our evaluation metrics are the same as the ones described for the synthetic data experiment in Sect. 4. The true mean, median, standard deviation and interquartile range of the complete smoothness feature are 0.096, 0.096, 0.014 and 0.018 respectively.

Next, we investigate implications of the drop method for classification accuracy. In this study, we apply random forests using the Scikit-learn library in Python with a maximum tree depth of 3 and no tuned hyperparameters. This random forest has an accuracy of 0.936 on the complete dataset. The incomplete dataset is analyzed using two scenarios as shown in Fig. 2:

a) An incomplete dataset is randomly split into training and test data, and incomplete records are dropped from both sets.[2]
b) The test data is a selection of complete records in an incomplete dataset. During development, incomplete cases will be dropped from the training data.

Although the first scenario is not applicable during deployment, it may still reveal interesting patterns since the smoothness feature is skewed with a factor 0.10. Consequently, different missingness types may affect the observed data distribution differently. The second scenario is generally not applied in practice (at least, we hope so), but provides an excellent way of studying the extent to which distribution shift affects classification accuracy.

[2] N.B.: in the general case, this may affect training and test distribution, but it is unclear how. Homogeneity might increase, but the data might also become more scattered and hence variance might increase. Since the distribution can be affected in a wide variety of possible ways, we will simply ignore this effect; note that technically this might affect the definition of accuracy.

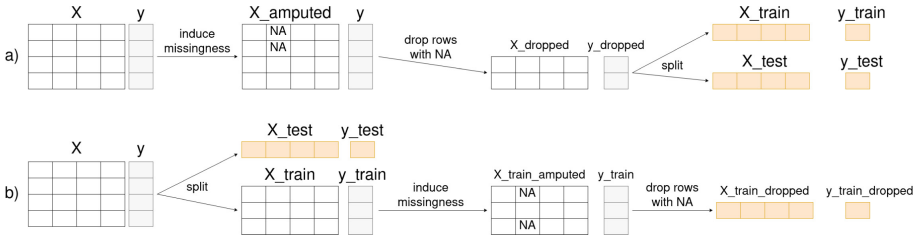

Fig. 2. Two scenarios for splitting incomplete data into train and test set.

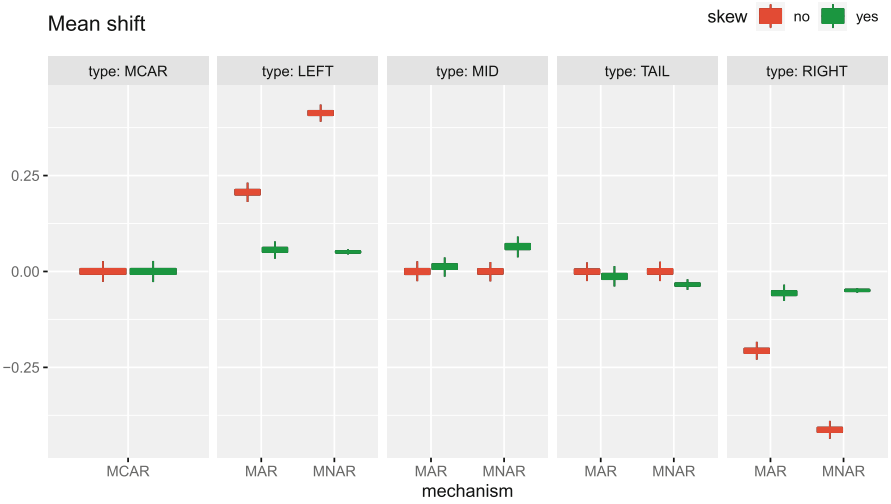

Fig. 3. Mean shift after dropping incomplete rows for 9 missing data scenarios.

5 Results

All results can also be found in our Github repository.

5.1 Results for Skewness

Figure 3 shows the mean shift for all 9 simulation settings; symmetrical data in red and skewed data in green. Without skew, results confirm existing knowledge that MCAR, and mid and tail types of MAR and MNAR missingness do not create mean shift. In contrast, left and right types of MAR and MNAR missingness shift the mean to the right (positive shift) and to the left (negative shift) respectively. MNAR generates more shift than MAR missingness.

When data has right-directed skewness, left and right missingness types create less mean shift than in the case of symmetrical data (compare the green and

Fig. 4. Standard deviation shift after dropping incomplete rows for 9 missing data scenarios.

red boxplots). Interestingly, for skewed data, the average mean shift is approximately similar for MAR and MNAR, although there is more variation across simulation repetitions for MAR. In addition, for skewed data, mid and tail missingness types induce mean shift such that mid missingness mimics left and tail missingness mimics right missingness.

Evaluating the shift in standard deviation for symmetrical data does not give unexpected results (see Fig. 4). Without skew, MCAR does not affect the spread of the distribution, while mid and tail missingness types respectively increase and decrease the standard deviation. Furthermore, for both left and right types of missingness the standard deviation is reduced.

Interestingly, for left missingness, right-directed skewness *increases* the standard deviation rather than decreasing it (compare green with red boxplots in second panel). Furthermore, when data is skewed, effects of missing data on the standard deviation vary more between simulation repetitions, especially for MCAR and MAR mechanisms. For measures of the median and interquartile range, results are similar as in Figs. 3 and 4 but less extreme.

5.2 Results for Mixtures of Mechanisms

We present the average mean shift over $T = 1\,000$ repetitions for mixtures of MCAR, MAR and MNAR mechanisms in Fig. 5. Naturally, for right missingness, we see that the higher the missingness percentage, the more the mean shifts. This increase is larger for MNAR than for MAR missingness. For instance, for 10% MNAR missingness (center of the figure), the increase per 10% of MAR missingness is around 0.0005 (from light orange to dark orange). In contrast,

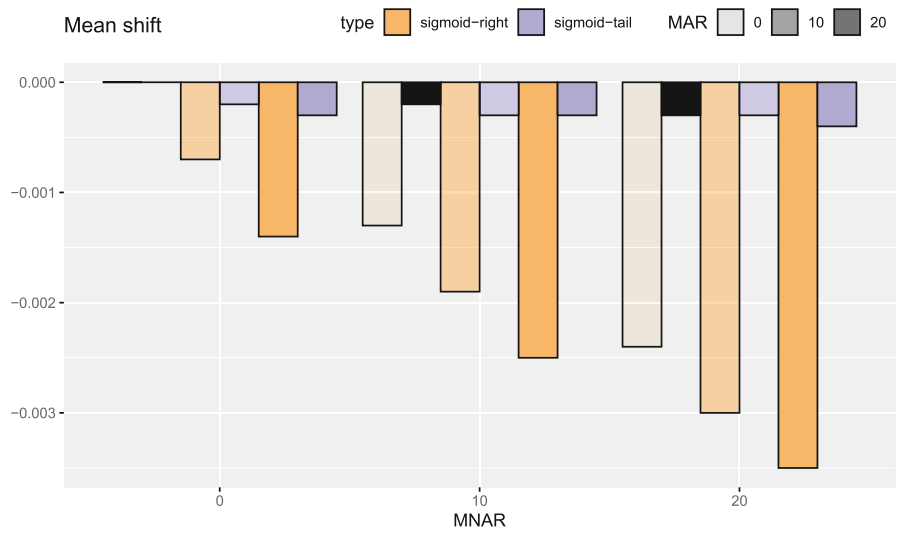

Fig. 5. Average mean shift for mixtures of MCAR, MAR and MNAR mechanisms. MCAR missingness is fixed to 10%.

for 10% MAR missingness (the very light orange bars), the increase per 10% of MNAR missingness is around 0.001. On average, the mean shift for MNAR missingness is twice the amount of the shift for MAR missingness.

Figure 5 demonstrates that the effect of combining multiple mechanisms is additive (rather than, for example, multiplicative). For instance, compare the mixture of 10% MNAR and 10% MNAR (medium orange, center of the figure) with a single MNAR mechanism of 20% (light orange, right side). It turns out that the former creates an approximate mean shift of $0.001 + 0.0005 = 0.0015$; the latter shifts $2 \cdot 0.001 = 0.002$. In addition, a higher missingness percentage may not necessarily result in more mean shift. For instance, the combination of 10% MNAR and 20% MAR (dark orange, center of the figure) shifts the mean with $0.001 + 2 \cdot 0.0005 = 0.002$, but a pure 20% MNAR mechanism has a similar effect. Our findings confirm that not only the missingness percentage determines the extent to which the mean shifts, but especially the occurrence of certain (combinations of) missingness mechanisms play a role.

Figure 6 displays the effects of left and mid missingness types on the standard deviation. Similarly as when we evaluated mean shift, combining multiple mechanisms has an additive effect. For instance, for mid missingness, 10% MNAR combined with 20% MAR increases the standard deviation with 0.0008. A pure 20% MNAR mechanism induces the same amount of shift.

Interestingly, a pure left-type of MNAR missingness *decreases* the standard deviation (light orange bars show negative shift). In contrast, combinations of MNAR and MAR missingness *increase* the standard deviation. Here, it seems that the MNAR component behaves as if the smoothness feature is symmetrically

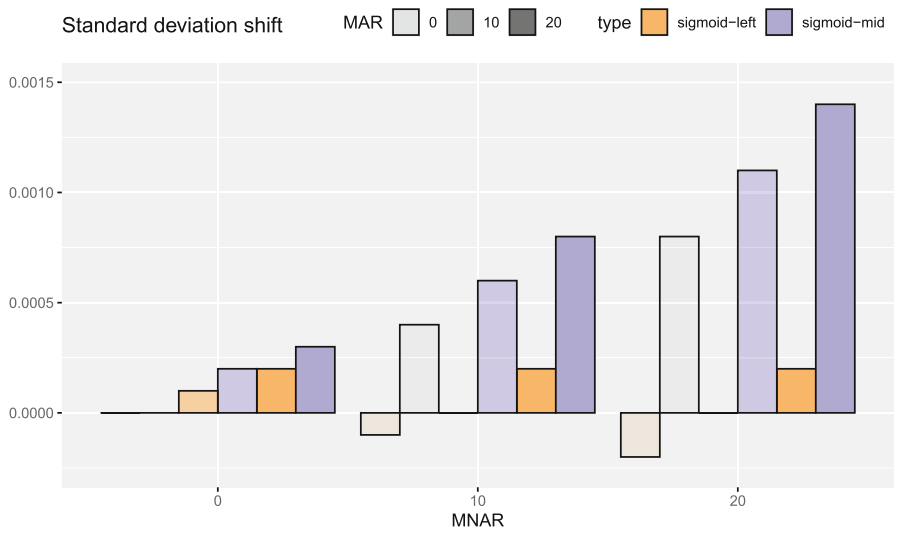

Fig. 6. Average standard deviation shift for mixtures of MCAR, MAR and MNAR mechanisms. MCAR missingness is fixed to 10%.

distributed, whereas the MAR component seems to be affected by the right-directed skewness in the feature (see Sect. 5.1). This may be explained by the fact that for MAR, missingness probabilities depend on observed data in the *symmetry* feature, which is skewed as well.

5.3 Results for Classification Accuracy

We present results for our investigation of the effect of the drop method on classification accuracy in Table 1. We present the correlation between the absolute shift and classification accuracy. Interestingly, when incomplete data is randomly split in training and test data and incomplete records are dropped in both datasets (split scenario a in Fig. 2), all significant correlations are positive (orange values), which means that a larger shift will increase the classification accuracy. This finding is rather counterintuitive, but there may possibly be an effect on the symmetry of the data such that records that are difficult to predict are dropped from both training and test data.

When the test data is a selection of complete records from an incomplete dataset (split scenario b in Fig. 2), all significant correlations are negative (teal values). Here, the larger the shift, the lower the classification accuracy. These findings confirm intuitions; if training data has a different data distribution than test data, classification accuracy will decrease.

Table 1. Correlation between the absolute mean and standard deviation shift and classification accuracy for two train-test split scenarios. MCAR = 0%, SEs are 0.011 in all scenarios, non-significant correlations are italicized. Orange and teal values present an increase and decrease in accuracy respectively.

sigmoid	**mean shift**		**std shift**	
	accuracy		**accuracy**	
	a	b	a	b
right	*0.009*	-0.165	*0.009*	-0.081
left	0.254	*0.006*	*0.017*	*0.014*
tail	*0.015*	-0.046	0.107	*-0.009*
mid	0.123	-0.034	0.191	-0.088

6 Discussion and Conclusion

Dropping incomplete records is straightforward; at least when there is no doubt about the effects on the center and spread of the data distribution. However, we demonstrated that when data contains right-directed skewness, tail and mid missingness types induce mean shift, left missingness increases rather than decreases the standard deviation, and the effects of MAR missingness fluctuate substantially. We furthermore showed that when multiple missing data mechanisms occur simultaneously, their effects on the data distribution are additive.

We evaluated the relation between dropping incomplete records and classification accuracy using a Random Forests (RFs) classification model. In reality, RFs are able to handle missing data by making surrogate splits or by treating missing values as a separate category. Moreover, [6] connect RFs to Probabilistic Circuits and propose Generative Forests (GeFs); a family of models that could handle incomplete features internally. Nevertheless, in this paper, our interest is not in creating the best classification model, but rather to show relations between data distribution shift and accuracy.

We found that classification accuracy decreases when training and test data are not identically distributed. Alternatively, when incomplete records are dropped before making a train-test split, accuracy increased. A possible explanation is that in such a situation, records that were difficult to predict were dropped. Note that classification accuracy changes when a record crosses the decision threshold; subtle differences in accuracy may be better detectable by evaluating a prediction model.

In sum, we showed that dropping incomplete records alters the data distribution considerably; some changes are straightforward, others are not. In general, our findings have implications for popular imputation methods such as mean imputation since imputations based on shifted data may transform the data structure even more.

6.1 Limitations and Future Work

This paper expands upon a long tradition of missing data research, historically driven by statisticians more than data miners. We explore the three canonical missing data mechanisms MCAR, MAR and MNAR as proposed by Rubin [15], and consider the four missing data types from [17] as also illustrated in Fig. 1. We believe that the conclusions we draw here are valid and well-supported by a rigorous set of experiments, but these experiments do come with some limitations. The experiments are run on variations of only a single dataset, apart from the class label all attributes are real-valued, the experiments employ only a single classifier, only a single feature has missing values, only unimodal distributions are investigated, further parameter sensitivity analyses could be imagined (for instance, do the conclusions change when the missingness rate is varied, or when missingness depends on the class label?). It is apparent that more work on this topic is to be done in the (near) future.

Acknowledgments. Many thanks to dr. Wouter Duivesteijn and prof. Mykola Pechenizkiy for their continuous support in all possible ways. Thank you Hilde Weerts for being a sparring partner.

References

1. Acuna, E., Rodriguez, C.: The treatment of missing values and its effect on classifier accuracy. In: Banks, D., McMorris, F.R., Arabie, P., Gaul, W. (eds.) Classification, Clustering, and Data Mining Applications. Studies in Classification, Data Analysis, and Knowledge Organisation, pp. 639–647. Springer, Berlin, Heidelberg (2004). https://doi.org/10.1007/978-3-642-17103-1_60
2. Brand, J.P., van Buuren, S., Groothuis-Oudshoorn, K., Gelsema, E.S.: A toolkit in SAS for the evaluation of multiple imputation methods. Stat. Neerl. **57**(1), 36–45 (2003)
3. van Buuren, S.: Flexible Imputation of Missing Data, 2nd edn. Chapman and Hall/CRC, Boca Raton (2018)
4. van Buuren, S., Brand, J.P., Groothuis-Oudshoorn, C.G., Rubin, D.B.: Fully conditional specification in multivariate imputation. J. Stat. Comput. Simul. **76**(12), 1049–1064 (2006)
5. van Buuren, S., Groothuis-Oudshoorn, K.: MICE: multivariate imputation by chained equations in R. J. Stat. Softw. **45**, 1–67 (2011)
6. Correia, A., Peharz, R., de Campos, C.P.: Joints in random forests. Adv. Neural Inf. Process. Syst. **33**, 11404–11415 (2020)
7. García-Laencina, P.J., Sancho-Gómez, J.L., Figueiras-Vidal, A.R.: Pattern classification with missing data: a review. Neural Comput. Appl. **19**(2), 263–282 (2010)
8. Garciarena, U., Santana, R.: An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. Expert Syst. Appl. **89**, 52–65 (2017)
9. Hoogland, J., et al.: Handling missing predictor values when validating and applying a prediction model to new patients. Stat. Med. **39**(25), 3591–3607 (2020)
10. Little, R.J.: Regression with missing X's: a review. J. Am. Stat. Assoc. **87**(420), 1227–1237 (1992)

11. Little, R.J., Rubin, D.B.: Statistical Analysis with Missing Data, Wiley Series in Probability and Statistics, vol. 793. Wiley, Hoboken (2019)
12. Mangasarian, O.L., Street, W.N., Wolberg, W.H.: Breast cancer diagnosis and prognosis via linear programming. Oper. Res. **43**(4), 570–577 (1995)
13. Miller, I., Miller, M., Freund, J.E.: John E. Freund's Mathematical Statistics, 6th edn. Prentice Hall, Upper Saddle River, N.J. (1999)
14. Raji, I.D., Kumar, I.E., Horowitz, A., Selbst, A.: The fallacy of AI functionality. In: ACM Conference on Fairness, Accountability, and Transparency, pp. 959–972 (2022)
15. Rubin, D.B.: Inference and missing data. Biometrika **63**(3), 581–592 (1976)
16. Schafer, J.L., Graham, J.W.: Missing data: our view of the state of the art. Psychol. Methods **7**(2), 147 (2002)
17. Schouten, R.M., Lugtig, P., Vink, G.: Generating missing values for simulation purposes: a multivariate amputation procedure. J. Stat. Comput. Simul. **88**(15), 2909–2930 (2018)
18. Schouten, R.M., Vink, G.: The dance of the mechanisms: how observed information influences the validity of missingness assumptions. Sociol. Methods Res. **50**(3), 1243–1258 (2021)
19. Schouten, R.M., Zamanzadeh, D., Singh, P.: pyampute: a python library for data amputation, August 2022. https://doi.org/10.25080/majora-212e5952-03e
20. Street, W.N., Wolberg, W.H., Mangasarian, O.L.: Nuclear feature extraction for breast tumor diagnosis. In: Acharya, R.S., Goldgof, D.B. (eds.) Biomedical Image Processing and Biomedical Visualization. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 1905, pp. 861–870, July 1993
21. Toutenburg, H., Srivastava, V.K.: Shalabh: amputation versus imputation of missing values through ratio method in sample surveys. Stat. Pap. **49**(2), 237–247 (2008)
22. Žliobaitė, I., Pechenizkiy, M., Gama, J.: An overview of concept drift applications. In: Japkowicz, N., Stefanowski, J. (eds.) Big Data Analysis: New Algorithms for a New Society. SBD, vol. 16, pp. 91–114. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-26989-4_4