Missing Values and Imputation in Healthcare Data: Can Interpretable Machine Learning Help?

Zhi Chen ZHI.CHEN1@DUKE.EDU
Duke University, USA

Sarah Tan HT395@CORNELL.EDU
Cornell University, USA

Urszula Chajewska URSZC@MICROSOFT.COM
Microsoft Research, USA

Cynthia Rudin CYNTHIA@CS.DUKE.EDU
Duke University, USA

Rich Caruana RCARUANA@MICROSOFT.COM
Microsoft Research, USA

Abstract

Missing values are a fundamental problem in data science. Many datasets have missing values that must be properly handled because the way missing values are treated can have large impact on the resulting machine learning model. In medical applications, the consequences may affect healthcare decisions. There are many methods in the literature for dealing with missing values, including state-of-the-art methods which often depend on black-box models for imputation. In this work, we show how recent advances in interpretable machine learning provide a new perspective for understanding and tackling the missing value problem. We propose methods based on high-accuracy glass-box Explainable Boosting Machines (EBMs) that can help users (1) gain new insights on missingness mechanisms and better understand the causes of missingness, and (2) detect – or even alleviate – potential risks introduced by imputation algorithms. Experiments on real-world medical datasets illustrate the effectiveness of the proposed methods.

Data and Code Availability: This paper uses two publicly available datasets: MIMIC-II (Saeed et al., 2002) and CDC Birth Cohort Linked Birth - Infant Death Data Files (United States Department of Health and Human Services (US DHHS) et al., 2021), and a proprietary pneumonia mortality prediction dataset (Cooper et al., 2005). The experiments leverage InterpretML open source (Nori et al., 2019) software package, and experiment code is provided in the supplementary materials.

Institutional Review Board (IRB): The research does not require IRB approval.

1. Introduction

Missing values are ubiquitous in most datasets and have significant impact on machine learning models, as most machine learning models do not naturally handle missing values. While one could simply delete rows or columns as a preprocessing step, so the learning algorithm is only given observed, non-missing samples as inputs, such methods only work when the missingness ratio is small and the feature values are missing completely at random (MCAR). Deleting cases with non-MCAR missing values risks changing the data distribution, losing what might be valuable information contained in the missing cases.

To avoid potential risks, systematic studies on understanding and handling missing values have been conducted in statistics and machine learning. Mechanisms of missingness have been studied and classified into three main categories, missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Different types of missingness have different solutions. For example, data cleaning and deletion methods like listwise and pairwise deletion are often used for MCAR (Rubin, 1976). In the MAR scenario, numerous imputation methods have been proposed. These include simple tech-

© 2023 Z. Chen, S. Tan, U. Chajewska, C. Rudin & R. Caruana.

niques like imputing missing values with the mean or median, using a unique value to denote missing, and using advanced statistical and machine learning models to impute the missing values. State-of-the-art imputation methods include discriminative models like MICE (Van Buuren and Groothuis-Oudshoorn, 2011), MissForest (Stekhoven and Bühlmann, 2011), KNN Imputer (Troyanskaya et al., 2001), and matrix completion (Mazumder et al., 2010; Yu et al., 2016), and generative models like deep generative models (Yoon et al., 2018). Since most of these methods are based on black-box machine learning methods and the accuracy and behavior of the final model depends on the imputed values, it is difficult for users to recognize and understand the potential harms that might be introduced by these imputation methods.

Recently developed interpretable machine learning methods have been shown to be useful for debugging models and detecting issues with datasets (Adebayo et al., 2020; Koh and Liang, 2017). Interpretable machine learning methods provide a new opportunity to study missing values and revisit some of the classical methods for handling missing values. In this paper, we propose novel methods based on the Explainable Boosting Machine (EBM) (Lou et al., 2012, 2013; Nori et al., 2019), a high-accuracy, fully-interpretable glass-box machine learning method, to answer the following questions: (1) how interpretability can help users gain insights on the causes of missingness, and (2) how interpretability can help detect and avoid potential risks introduced by different imputation methods. We show that the glass-box models provide new insights into missingness mechanisms, and in some settings, suggest alternate ways of handling missing values, as well as new tools that can alert users when imputation can lead to unexpected problems.

2. Related Work

Issues with missing value imputation methods, whether generative or discriminative, have been pointed out in the literature (Harel and Zhou, 2007; Jeličić et al., 2009; Ibrahim et al., 2012; Li et al., 2015; Van Buuren, 2018; Sidi and Harel, 2018). For example, generative imputation methods have been criticized for placing assumptions on the underlying data distribution, not all of which are testable (Yoon et al., 2018). Waljee et al. (2013) studied four discriminative imputation methods – MissForest, mean imputation, nearest neighbor imputation, and multivariate imputation by chained equations (MICE) – on medical datasets modified to have missing completely at random (MCAR) values, finding that MissForest had the least imputation error for both continuous and categorical variables. Our paper shows that MissForest, despite its popularity, presents issues that practitioners should notice.

Connections between missing value imputation and causal inference methods have been drawn. Ding and Li (2018) pointed out that the unconfoundedness assumption in causal inference is similar to the missing at random (MAR) assumption in missing data analysis, with both fields relying on these untestable, yet critical assumptions. The interpretability techniques we use in this paper can be applied to datasets even if they have missing not at random (MNAR) values. This flexibility presents a contribution given how difficult it is to distinguish between MAR and MNAR in practice (Van Buuren, 2018).

Our work is related to recent work using explainability techniques to detect issues with datasets. Adebayo et al. (2020) investigate the ability of feature attribution methods to detect spurious correlations and mislabeled examples. Koh and Liang (2017) used influence functions applied to black-box models to detect mislabeled examples in data. Our work does not use black-box models, and focuses on debugging missing values, a key issue in many datasets.

Some AutoML tools perform automatic data cleaning. Both the Automatic Statistician (Steinruecken et al., 2019) and AlphaClean (Krishnan and Wu, 2019) attempt to automatically impute missing values. Unlike these papers, our focus is not on fixing datasets automatically, but on helping users detect, understand and mitigate missing values problems.

3. Background

3.1. Types of Missing Values

Rubin (1976) classified missingness mechanisms into three types: (1) Missing Completely At Random (MCAR): the missingness is unrelated to the data, i.e. the probability of missing is the same for all samples; (2) Missing At Random (MAR): in addition to complete randomness, the probability of missingness of a feature is determined from the observed values of the other features (3) Missing Not At Random (MNAR): the probability of missingness is also related to unobserved values in the data, e.g., the missingness is also related to the feature value itself.

3.2. Missing Value Imputation

Here, we describe the advanced imputation methods we investigate in this paper: MissForest (Stekhoven and Bühlmann, 2011) and KNN Imputation (Troyanskaya et al., 2001).

The MissForest algorithm first makes an initial guess for the missing values using mean and mode imputation. Then it sorts the features according to the missing rate, and fits a random forest iteratively to predict and impute each missing feature from the other features until the imputed values converge. MissForest is a popular imputation method as it is capable of capturing non-linear and interaction effects between features to improve imputation accuracy, and can be applied to mixed data types (continuous and discrete). Note that, the framework of MissForest is similar to that of MICE (Van Buuren and Groothuis-Oudshoorn, 2011) — the only difference is MissForest uses random forest while MICE uses linear model as base model for imputation.

KNN imputation imputes the missing values by the mean value of its K nearest neighbors in the training set. The distance of two samples is measured on the non-missing features in both samples. KNN imputation is fast and accurate but requires choosing a good distance metric and tuning the hyperparameter K.

3.3. Explainable Boosting Machines

The methods proposed in this work are based on one interpretable machine learning model, the Explainable Boosting Machine (EBM).

Suppose an input sample is denoted as $(\mathbf{x}, y)$, where $\mathbf{x}$ is the p dimensional feature vector and y is the target. Denote the j^{th} dimension of the feature vector as x_j. Then a generalized additive model (GAM), first introduced by Hastie and Tibshirani (1987), is defined as

$$g(E[y]) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p) \quad (1)$$

where β_0 is the intercept, $f_j's$ are the shape functions and g is the link function, e.g., the identity function for regression, or the logistic function for classification. Since one can add any offset to f_j while subtracting it from β_0 or other shape functions, shape functions are often *centered* by setting the population mean of f_j, i.e., $E_{x\sim\mathcal{X}}[f_j(x_j)]$ to 0. Because each shape function f_j operates only on one single feature x_j, shape functions can be plotted. This makes GAMs interpretable since the model can be visualized as 2D graphs. In early work on GAMs, shape functions were often modeled as splines with smoothness constraints. Explainable Boosting Machines (EBMs) (Lou et al., 2012, 2013; Nori et al., 2019) use bagged ensembles of boosted depth-restricted tree to represent each f_j. Tree-based ensemble learning significantly improves the performance of GAMs: EBMs outperform traditional GAMs because its shape functions have more representational power and better capture fine detail. Figure 1 shows the shape plot learned for P/F ratio (a measure of blood oxygenation) on the MIMIC II ICU mortality-risk classification problem. The vertical axis is the contribution to risk on log scale: patients with low P/F ratio have high risk, and patients with P/F ratio near 1000 are low risk. EBM can further improves its accuracy by adding a small number of pairwise interactions, i.e.,

$$g(E[y]) = \beta_0 + \sum_{j=1}^{p} f_j(x_j) + \sum_{k=1}^{K} f_k(x_{k_1}, x_{k_2}). \quad (2)$$

Including pairwise interactions does not sacrifice interpretability since $f_k(x_{k_1}, x_{k_2})$ can be visualized as heatmaps. In this paper, we use EBM implemented in the InterpretML package (Nori et al., 2019).

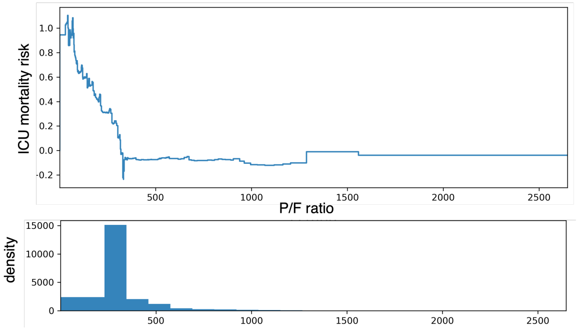


Figure 1: EBM shape function and density plot for P/F ratio when predicting ICU mortality.

4. Gaining new insights on the causes of missingness

4.1. Missing Completely at Random

When dealing with missing values, it is important to determine the mechanism of missingness. Standard statistical tests exist for testing MCAR, e.g., Little's test (Little, 1988). In this section, we propose a method to test MCAR based on EBM shape functions. The testing process of the proposed method can be directly visualized on the shape function plots, which is not achievable by Little's test. We will also show that EBM can bring additional insights beyond simply testing for MCAR.

4.1.1. Testing for MCAR with EBM

To test for MCAR, we use the common trick of encoding missing values with a unique value for the feature, e.g., -1 for a feature with positive values or a separate category for a categorical feature. After fitting an EBM that predicts the target, we get a shape function representing the contribution of different feature values for predicting the target, including the unique value denoting missingness. Note that the leaf nodes in EBM split the feature values into many bins, where each bin has a prediction score. These bins and scores together form the shape function. Therefore, the EBM shape function $f_j(\cdot)$ of feature j can be rewritten as a linear combination of a series of indicator variables denoting if the feature values are within the bins, and the coefficients are the corresponding scores of the bins, i.e.,

$$f_j(x_j) = \sum_{k=0}^{B_j-1} \theta_{j,k} \cdot \mathbb{1}\{b_{j,k} < x_j <= b_{j,(k+1)}\}, \tag{3}$$

where $\{b_{j,k}\}_{k=0}^{B_j}$ are the bin edges of feature j in the EBM model, and $\theta_{j,k}$ is the shape function score of the bin $(b_{j,k}, b_{j,(k+1)}]$. Since EBM also uses the logistic link function, this transformation can turn EBM into a logistic regression model for binary classification. To do a statistical test, we need to make some assumptions and create a null hypothesis. First, we know that if the missingness is MCAR, i.e., all samples are missing with the same probability, the expected score for bins representing the missing value should be the same as the entire population, which is 0 as the shape function scores are mean centered in EBMs. Therefore, we can directly apply the classical significance Wald test of logistic regression coefficients (Kleinbaum et al., 2002). Specifically, our null hypothesis is $H_0: \ \theta_{i,k} = 0$, and the alternative hypothesis is $H_1: \ \theta_{i,k} \neq 0$. Then we calculate the square root of the Wald statistic

$$\sqrt{W} = \frac{\hat{\theta}_{j,k}}{SE(\hat{\theta}_{j,k})}, \tag{4}$$

find the p-value by assuming $\sqrt{W}$ follows a Z distribution, and reject the null hypothesis if the p-value is smaller than a predefined threshold. Figure 2 shows an example of using the proposed test for MCAR using EBM shape functions. The missing value is encoded as -5 (lower than minimum possible feature value) for the Bilirubin feature. The Wald test rejects the null hypothesis and suggests that the missing value is not MCAR.

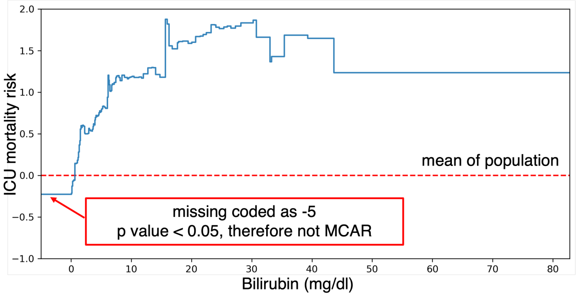


Figure 2: Example of using EBM shape function to test for MCAR. Missing coded as -5. p-value for testing MCAR is less than 0.05.

Type	MCAR datasets↓			MAR datasets↑		
p_m	0.1	0.2	0.3	0.1	0.2	0.3
Little's	**0.035**	0.070	0.055	**1.000**	**1.000**	**1.000**
Ours	0.080	**0.005**	**0.005**	0.910	0.885	0.890

Table 1: Proportion of times the test rejects the null hypothesis, i.e., the missing mechanism is MCAR, on datasets generated by different missing types, with different missing ratios p_m. When the data is truly MCAR, for which low rejection rate is desired, our method is less likely to reject the null hypothesis compared with Little's test, especially when missing ratio becomes larger. When the data is MAR, where we hope to reject the null hypothesis (high rejection rate is better), Little's test can reject all null hypothesis, and our method is able to reject it in most cases.

Table 1 compares the performance of the MCAR test we proposed with Little (1988). To test their performances, we generate semi-synthetic datasets where we know the ground truth missing mechanism. Specifically, we start from MIMIC-II dataset imputed by MissForest, and then add missing values to the "Age" feature manually[1]. For MCAR case, each sample has a fixed probability p_m of missing the "Age" feature. For MAR case, we apply a linear model on all features except "Age", whose coefficients are ran-

[1] The "Age" feature has no missing values in the original dataset.

domly sampled from standard normal distribution to all samples in the dataset, adding a standard Gaussian noise to the output score, and the $\lceil np_m \rceil$ samples with the lowest output scores from the linear model (plus noise) are missing the "Age" feature. We generate 200 datasets with MCAR values and 200 datasets with MAR values, and apply both our MCAR test and Little's test to these datasets, and check if these test will reject the hypothesis that the missing is MCAR (p value<0.05). Table 1 shows the ratio of rejecting null hypothesis. Our method detects MCAR values more reliably than Little's method in high (20%-30%) missingness cases.

Summary: We propose an application of EBM to test if the missing value is MCAR. See Appendix A for a case study on infant mortality risk showing that such method may be useful in determining applicability of a model for future data.

4.2. Missing Values Assumed Normal

In healthcare domain, it is common for feature values such as lab tests to be missing in the dataset because clinicians believed the patient was likely to be "normal" for this measurement, and thus the lab test was not performed (Li et al., 2021). In other cases, the measurement may have been made, but the value was not recorded since it was within normal range — clinicians tend to focus on abnormal findings.

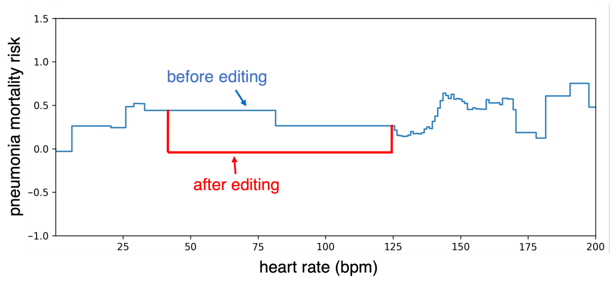


Figure 3: EBM shape function of "heart rate" for predicting pneumonia mortality risk. Blue curve is the original shape function; red curve is the edited shape function.

For example, this happens to a pneumonia mortality risk dataset (Cooper et al., 2005). The blue curve in Figure 3 shows what an EBM model has learned for predicting pneumonia mortality as a function of heart rate. As expected, risk is elevated for patients with abnormally low (10-30) or high heart rate (125-200). The graph, however, shows a surprising region of flat risk between heart rate 38 and 125, which is a normal heart rate for patients in a doctor's office. Moreover, the model surprisingly predicts patients who have *normal* heart rate are at *elevated* risk: it adds 0.22 to the risk for patients in this region.

On further inspection, it turns out that there are no patients in the data set with heart rates between 38 and 125, and 91% of patients are missing their heart rate which has then been coded as zero. In other words, there are no data to support the model in the normal range of heart rate 38-125, and instead the patients who would be in this range are all coded as zero in the data and on the graph. This explains why the model predicts the lowest risk = -0.04 for patients with heart rate = 0, because these are the patients who actually have normal heart rates.

Any model trained on this data (e.g., boosted trees, random forest, neural networks) is likely to learn to make similar predictions as EBMs in the normal heart rate region because there is no data to support learning the correct risk in this range, and because most models will then interpolate between the extreme regions where they do have data. One exception might be Bayesian models with strong priors, where the prior might dominate in regions of little or no data and cause predictions in this region to be closer to a baseline lower-risk value. The key advantage of using interpretable models such as EBMs is that we can easily see these problems in the model, that ultimately were caused by problems in the data.

If patients with normal heart rates (38-125) will always be coded as zero in the future, then a model trained on this data might always make accurate predictions and the elevated risk predicted by the model in the range 38-125 will not be a problem because no patient will ever fall in that range. However, if the model might be used to make predictions for patients whose true heart rate would be coded between 38 and 125, the model will then make incorrect – possibly dangerous – predictions. Thus, it is important to correct this kind of problem. One might hope that a data scientist would detect this kind of problem in the data prior to training a model, however in practice, these kinds of problems can be difficult to detect in the raw data, particularly if there are many different types of problems in the data, and might be easier to detect once an interpretable model is trained. (Previous users of the data had not noticed this problem.)

4.2.1. Correction via Model Editing

There are several ways to correct this kind of problem. Of course, the best approach would be to collect and record the true heart rates for all patients. Unfortunately, it is often not possible to go back and correct data in this way. As we will see in Section 5.1, imputing with the mean or median missing value is probably not ideal. We will also show in Section 5.2 that more advanced methods of imputing missing values such as random forest imputation (Stekhoven and Bühlmann, 2011) and KNN imputation (Troyanskaya et al., 2001) might also cause problems.

An alternate approach when EBMs are used is to directly edit the model so that the region 38-125 predicts risk similar to the learned risk prediction for patients with heart rate = 0. Since we do not have any information about true heart rate distribution within the region 38-125, we assume they follow a uniform distribution and edit the graph in this region to be a flat curve. The resulting graph is shown as the red curve in Figure 3. (Note that the result would be similar to uniformly imputing the heart rates in the interval 38-125 and retraining the model.) This approach has the following advantages:

1. Editing shape functions provides an opportunity for experts to use their professional training to correct and improve models in ways that may not be adequately represented in the training data.

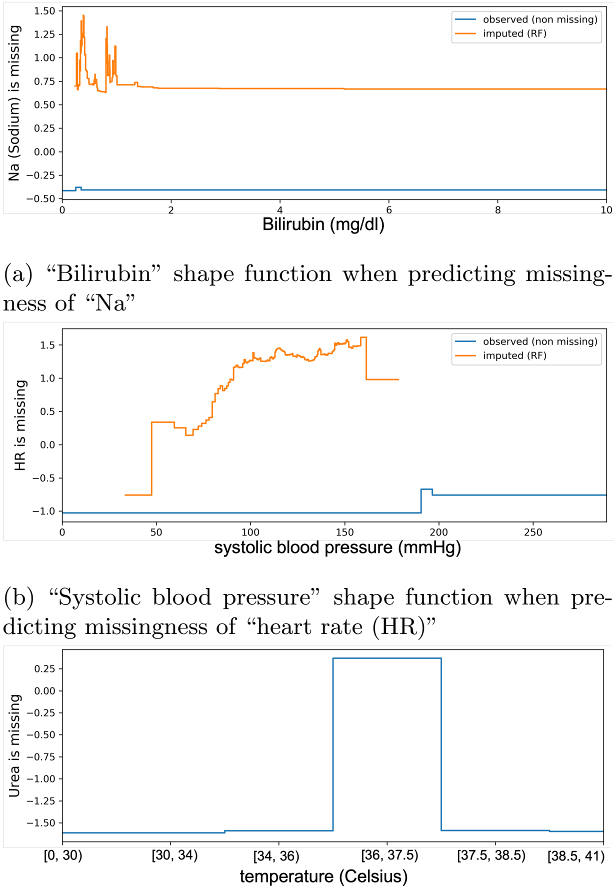
2. Editing the model may not only improve the accuracy of the model in the real world where it will be used (instead of just on held-aside test data from the train set), but also make the shape plots more “reasonable” and trusted by experts.

3. Editing an EBM shape function can be done without retraining the model and potentially introducing new problems.

Summary: We show that EBM shape function can help identify the case when feature values are missing because they are assumed to be normal. We also show how editing the EBM graphs can help address issues resulting from missing assumed normal.

4.3. Predicting the Missingness

Most missing values are not MCAR, but as mentioned in Section 2, MNAR and MAR can be difficult to distinguish (Van Buuren, 2018). For both cases, interpretable models like EBM can still be useful in providing insights on possible missingness mechanisms.



(a) “Bilirubin” shape function when predicting missingness of “Na”

(b) “Systolic blood pressure” shape function when predicting missingness of “heart rate (HR)”

(c) “Temperature” shape function when predicting missingness of “Urea”

Figure 4: EBM shape functions for predicting the missingness of one feature using the others (x-axis: feature value, y-axis: contribution to missingness). The effects of the imputed group (orange) and the non-missing group (blue) are separated.

One way to analyze the missingness mechanism is to predict the missingness of one variable using the other variables (including the target/label). Specifically, the 0-1 missingness indicator is considered as label, and the other features and the label of the original prediction task are considered as input feature to train the machine learning model. The prediction accuracy for missingness tells us roughly how much the missingness is related to the values of other variables. More importantly, with the interpretability of EBMs, we can visualize how the values of these variables contribute to the missingness.

model	p_m	linear	curvilinear	quadratic
LR	0.1	0.954±0.014	0.902±0.016	**0.883±0.02**
RF		0.943±0.014	0.946±0.013	**0.883±0.02**
KNN		0.895±0.013	0.894±0.009	0.881±0.021
EBM		**0.956±0.015**	**0.959±0.013**	0.881±0.02
LR	0.2	0.928±0.019	0.839±0.034	0.815±0.013
RF		0.911±0.019	0.928±0.019	**0.831±0.017**
KNN		0.813±0.024	0.81±0.022	0.812±0.008
EBM		**0.930±0.019**	**0.946±0.02**	0.822±0.016
LR	0.3	0.906±0.022	0.809±0.054	0.710±0.025
RF		0.887±0.021	0.926±0.019	**0.812±0.03**
KNN		0.744±0.032	0.752±0.042	0.711±0.016
EBM		**0.908±0.022**	**0.946±0.02**	0.795±0.03

(a) datasets generated by MAR

model	p_m	linear	curvilinear	quadratic
LR	0.1	0.957±0.013	0.901±0.013	**0.886±0.017**
RF		0.944±0.013	0.948±0.011	**0.886±0.017**
KNN		0.899±0.012	0.898±0.01	0.885±0.018
EBM		**0.959±0.012**	**0.963±0.011**	0.885±0.017
LR	0.2	0.928±0.018	0.847±0.035	0.817±0.010
RF		0.910±0.016	0.933±0.016	**0.828±0.012**
KNN		0.816±0.024	0.82±0.025	0.813±0.008
EBM		**0.931±0.017**	**0.953±0.016**	0.819±0.012
LR	0.3	0.914±0.016	0.805±0.048	0.706±0.024
RF		0.891±0.015	0.925±0.015	**0.811±0.028**
KNN		0.760±0.035	0.764±0.039	0.711±0.017
EBM		**0.916±0.016**	**0.949±0.015**	0.789±0.03

(b) datasets generated by MNAR

Table 2: Test accuracy of predicting the missingness. EBM is compared to Logistic Regression (LR), Random Forest (RF), and K Nearest Neighbor(KNN). The accuracies are compared on datasets generated by different missing mechanism (MAR and MNAR generated from linear model, curvilinear model, and quadratic model) with different missing ratio p_m (0.1, 0.2, and 0.3).

We train EBMs to predict missingness on the MIMIC-II dataset (Saeed et al., 2002) for every feature that contains missing values. The test AUCs for missingness prediction ranges from 69.10% to 99.87% depending on the missing feature: the test AUC is above 84% for 7 of the 9 missing features. Surprisingly, the test AUC for predicting missingness of "Na (Sodium)" and "Urea" are 98% and 99%, which suggests their missingness can be almost fully explained by other observed variables. Figure 4 shows the shape functions on MIMIC-II, which result from training an EBM on all other variables to predict the missingness of one variable. The features shown in Figure 4 are the features with the largest variable importance for each prediction task. Each shape function shows the contribution of the feature (on the x axis) to the predicted missingness (on the y-axis). Interesting patterns exist in all three graphs and provide insight about why each variables is missing.

Figure 4(a) shows how bilirubin contributes to predict the missingness of Na (Sodium). Though bilirubin is a continuous variable and we might expect the shape function to be a continuous curve, the shape function of the observed (non missing) bilirubin samples (in blue) is a constant function with contribution -0.4. This suggests that when bilirubin is measured, Na is less likely to be missing. Moreover, when bilirubin is missing and imputed (in orange), there is a large positive contribution (average contribution = +0.93) to the likelihood of Na missingness, which suggests that missing bilirubin strongly predicts that Na will be missing, too. Interestingly, the causal arrow does not flow the other way: Na is not a strong predictor of bilirubin missingness. The AUC when predicting Na missingness is 0.99, but only 0.73 for predicting bilirubin missingness, and the most important feature for predicting bilirubin missingness is Urea, not Na. All of this makes clinical sense because bilirubin is included in comprehensive metabolic panels that also always include Na, whereas basic metabolic panels include Na but not bilirubin, which is a more specialized lab test. This also explains why the non-missing group shape function (blue curve) is constant: patients whose Bilirubin are not missing took the comprehensive panels and thus their Na is always measured regardless of the patients' bilirubin value. Remarkably, we are able to detect and understand these effects merely by looking at interpretable EBM models trained to predict missingness.

We see a similar relationship between heart rate (HR) and blood pressure: when blood pressure is measured, heart rate is almost always measured as well, but it is common to measure heart rate using a finger sensor that does not allow blood pressure to be measured, and this asymmetric relationship between missingness is easily visible by examining EBM plots trained to predict HR missingness. Figure 4(b) shows the shape functions for observed systolic blood pressure (in blue) and imputed systolic blood pressure (in orange) when predicting whether HR is missing. In the plot, the curve of the imputed group is significantly higher than that of the observed group, again suggesting that when the blood pressure of the patients is missing, their heart rate is also more likely to be missing. This effect is strong, as the maximum gap between the two curves is approximately 2.5 (1.5 in orange curve and -1.0 in blue curve) of predicted log odds. Again the blue curve is constant.

Figure 4(c) shows the shape function for temperature when predicting if urea is missing or not. There is no missing value for temperature, so there is no orange curve. The bump at temperature $\in [36, 37.5)$ indicates that urea is more likely to be missing, which suggests when a patient has normal body temperature, doctors may be less likely to order a blood test to measure urea.

To test how well can EBM predict the missingness, we generate some semi-synthetic datasets with ground-truth missing mechanism. Again, these semi-synthetic datasets start from MIMIC-II imputed by MissForest, and then apply fixed models (linear, curvilinear and quadratic models) plus an Gaussian noise to decide which entry in the "Age" feature is missing. The feature value is missing when the output score is higher than the threshold. The difference between MAR and MNAR is whether the target feature value is considered as an input of the missing models. Table 2 compares EBM's the test accuracy of predicting missingness with machine learning models commonly used for missing value imputation. EBM predicts missingness better than other methods in cases of MAR and MNAR values generated from linear and curvilinear models and is not far behind Random Forest in case of quadratic model.

Summary: We use EBMs to predict the missingness of features from other input features. EBM predicts the missingness accurately. The interpretability of EBMs can help users understand the relationship between the features and missingness and thus bring more insight for the cause(s) of missingness.

5. Detecting and avoiding potential risks of missing value imputations

5.1. Imputation With the Mean

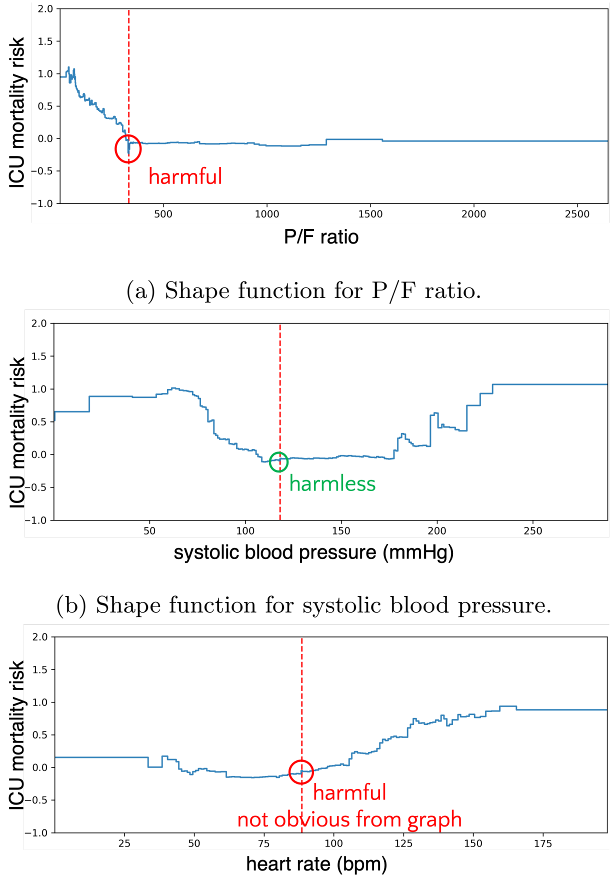
Because many machine learning methods cannot natively handle missing values, it is common for data scientists to impute missing values before training models. There are many different ways to do this (Lin and Tsai, 2020): with the mean, the median, with a unique value such as 0 or -99 or +99, or by using a machine learning method such as MissForest.

Perhaps the most common form of missing value imputation is to use the mean, but this can sometimes be problematic. Figure 1 shows an EBM plot of the mortality risk of ICU patients as a function of their P/F ratio. P/F ratio is a measure of how well a patient converts oxygen in the air they breathe into oxygen in their blood: low P/F ratio indicates patients with low blood-oxygen whose lung function is impaired, while P/F ratio around 1000 and higher indicates good lung function. As expected, the learned shape function captures this. What is surprising, however, is the large drop in risk at about P/F ratio=323. What could cause that?

A simple test for blood-oxygen levels is to pinch a fingertip and see how quickly color returns to the skin. If color returns quickly, clinicians know the blood-oxygen level is good and do not bother to measure P/F ratio — the P/F ratio is assumed normal. In this dataset, however, the missing P/F ratio values were imputed with the mean instead of being coded as 0 as they were in Figure 3. 60% of patients are missing P/F ratio. The mean P/F ratio when not missing (40% of the data) is 323.6, so 60% of patients have had their P/F ratio imputed with this value. Because this is a large sample of healthy patients with strong respiration, the model learns that their risk is comparable to the risk of other healthy patients with P/F ratio above 1000. This explains why the graph dips at 323, yet predicts higher risk just before and after this value. Although this anomaly does not significantly hurt the accuracy of the model because it has learned to make appropriate low-risk predictions for the 60% of patients at this value, it is risky to leave this anomaly in the model because there are real patients with P/F ratio≈323 who will be predicted to have low risk but who are genuinely at elevated risk. For this reason, it would be better to encode the missing value with unique value (e.g., -1). Model editing is not a good solution for this problem because imputation with the mean has caused patients who are low risk (missing values) and elevated risk (P/F ratio near 323) to fall at the same place on the shape function, thus there is no reasonable edit to the graph that can predict the correct risk for both groups.

5.1.1. Automatic Detection of Bad Imputations

As discussed in the P/F ratio example above, mean imputation could be dangerous especially when the missing group is significantly different from the samples with feature values near the mean. As shown in Figure 1 and Figure 5(a), such distribution differences can be reflected as spikes on the EBM shape functions. However, if the spike is small or there is no spike near the mean value, e.g., Figure 5 (b), the difference between groups might be insignificant and mean imputation can be harmless. Since bad mean imputation is associated with the spikes at the mean, can we automatically detect bad mean imputations through a spike detector? The answer is yes, but we need to address two problems, (1) how to know the spike is at the mean (2) how to detect spikes, given that the shape function itself can fluctuate.



(a) Shape function for P/F ratio.

(b) Shape function for systolic blood pressure.

(c) Shape function for heart rate.

Figure 5: Examples of potentially harmful (a) & (c) and harmless mean imputations (b) found by our automatic detection algorithm. Red vertical lines indicate average feature values. A large spike at the mean is potentially harmful while a small spike or no spike is harmless.

The first problem is easy to solve. Observing that the mean value of the feature is the same before and after mean imputation, we can directly find the bin (of EBM) covering the mean value, and detect if the bin is a spike or not. This also works for median imputation — the median of a feature does not change by imputing the missing values with the median.

To address the second problem, we need an algorithm to distinguish spikes resulting from mean imputation and fluctuations that naturally occur in the EBM shape functions. We formulate this as an outlier detection problem. First, we calculate the second order differences for all bins in all shape functions (excluding first and last bins), since spikes usually have extreme second order differences. We denote the function values of the k^{th} bin and its neighbouring bins as f_k, f_{k-1}, and f_{k+1}. The corresponding bin sizes are denoted as h_k, h_{k-1}, and h_{k+1}. The second order difference is

$$f''_k(x) \approx \frac{\frac{f_{k+1}-f_k}{(h_{k+1}+h_k)/2} - \frac{f_k-f_{k-1}}{(h_k+h_{k-1})/2}}{h_k + h_{k+1}/2 + h_{k-1}/2}. \tag{5}$$

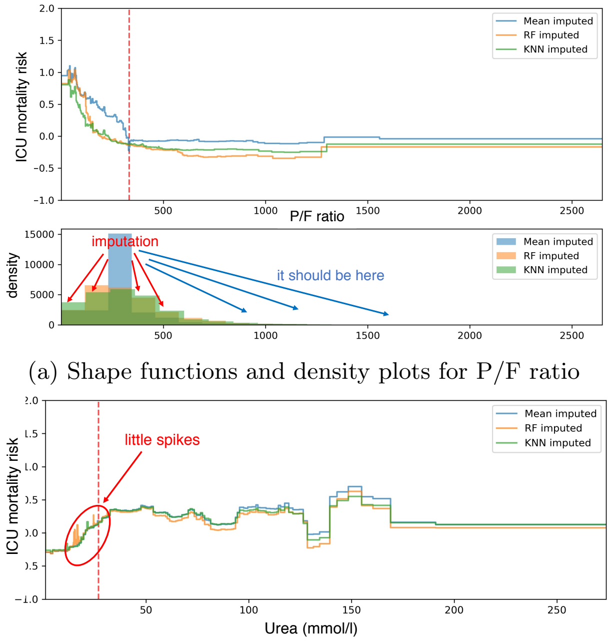
We then run an outlier detection algorithm (Isolation Forest (Liu et al., 2008)) on these second order differences. The algorithm predicts an anomaly score for each bin, and we choose a threshold so that around 5% of bins are detected as outliers. The potentially harmful mean imputations are predicted if bins covering the mean values are also predicted as outliers. The same procedure is also applied to detect potentially harmful median imputations.

We test the bad mean imputation detection algorithm on the MIMIC-II dataset with mean imputation on continuous features. Among the 13 continuous features, in 4 a spike is detected at the mean. Other continuous features do not have a spike at the mean and are predicted to be "harmless" in terms of mean imputation. As expected, continuous features with *no* missing values are predicted as negative. Figure 5(c) shows a potentially harmful mean imputation found by our detection algorithm but not discovered visually as the spike is not obvious. This represents one of the smallest spikes that the anomaly detection algorithm would detect as potentially harmful mean imputation (given this sample size).

Summary: By examining anomalies in the EBM shape functions one can easily identify bad imputations with the mean. Based on this finding, we propose an automatic detection method to detect imputations with the mean that can be potentially risky.

5.2. Imputation With Advanced Methods

One might assume that imputation with more advanced methods such as MissForest (RF) imputation (Stekhoven and Bühlmann, 2011) or k-nearest neighbor (KNN) imputation (Troyanskaya et al., 2001) would not exhibit problems like those discussed in



(a) Shape functions and density plots for P/F ratio

(b) Shape functions for Urea

Figure 6: EBM shape functions trained on datasets imputed with different missing value imputation methods (mean imputation, MissForest (RF) imputation and KNN imputation).

Sections 4.2 and 5.1, because they are designed to impute feature values based on the conditional feature distribution in the data. For example, MissForest iteratively trains a random forest regression model, predicting and updating the missing values of each covariate using the other covariates, until these values converge. Interpretable EBM models can help us detect unexpected problems that can be caused by imputation with these methods.

We apply different imputation methods (mean imputation, MissForest (RF) and KNN imputation) to the MIMIC-II dataset. Figure 6(a) shows the P/F ratio shape functions and densities for the different imputation methods. As described in Section 5.1, values near 1000 are healthy, and lower P/F ratio indicates poor lung function. In this dataset, P/F ratios are missing when doctors assume they are normal, i.e., the ground truth of missing values are likely to be near 1000. However, the density plot shows that instead of imputing missing values with P/F ratio values near 1000, RF and KNN actually impute P/F ratio with lower values. Such imputations are problematic because they systematically reduce the predicted risk of the riskier low-P/F-ratio patients and those patients might then not receive adequate care if the resulting model is used clinically. Compared with mean imputation, the advanced imputation methods actually affect a larger range of patients (P/F ratio between 0 and 800) and the advanced methods could be even more harmful than mean imputation.

Another problem of advanced imputation methods is that they can sometimes introduce fluctuations to models which show up as little spikes on EBM shape functions. Figure 6(b) shows the EBM shape function for the feature “Urea” with many little spikes when the missing values are imputed with RF and KNN. Again, such spikes can be potentially harmful for patients with almost the same feature values at these locations. The fluctuation problem can be resolved if the model enforces local smoothness (e.g., linear models or GAMs with smooth splines). However, tree-based models like random forests, gradient boosted trees, and EBMs often are not locally smooth and are likely to learn such spikes.

Summary: We show that advanced imputation methods like MissForest and KNN can create problems for machine learning models that are hard to detect. We propose a way to use EBMs to visualize the potential impact of these imputation methods (Appendix B), and show that it helps detect potential problems that otherwise might have remained invisible and led to suboptimal healthcare decisions.

6. Discussion

We found many potential risks in models that were introduced by missing values or imputation. Because EBMs are interpretable and editable, once the problem is detected, we can often edit the model to fix these issues using existing model editing tools for GAMs (Wang et al., 2021). Because edits only affect model behavior on small subsets of samples and for a few features (e.g., samples near the mean in the case of mean imputation), the change in accuracy is small. However, these changes can still be critical in high-stakes tasks like medical care, where the potential cost for bad predictions is very high.

The proposed methods are all based on EBM. We chose EBMs because the shape functions are good at capturing subtle anomalies in the data, compared to linear models and decision trees. In the future, it is worth investigating if other interpretability methods can handle the same missing value tasks. For example, a sparse decision tree model (Lin et al., 2020) might be able to learn complex feature interactions when predicting missingness from other features.

7. Conclusion

We propose methods based on glass-box EBMs to help understand and address missing value problems. Such problems are common in medical applications. Experiments on real-world medical datasets show that the proposed methods provide insights on the causes of missingness, and can also help detect and avoid potential risks introduced by different imputation methods. Specifically, in terms of understanding missingness, we propose a novel method using EBMs to test for MCAR. For the non-MCAR case, we show that EBM shape functions can help identify when feature values are missing because they were assumed to be in the normal range for that variable. We also use EBMs to predict the missingness of some features from other input features. Here the interpretability of the model can help users better understand the relationship between features and missingness. For imputation, we show that anomalies in the EBM shape functions can be used to automatically identify potentially harmful imputation with the mean or median. For advanced methods like MissForest and KNN imputation, we propose methods for visualizing the potential impact of imputation on the resulting model.

References

Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. In *NeurIPS*, 2020.

Gregory F Cooper, Vijoy Abraham, Constantin F Aliferis, John M Aronis, Bruce G Buchanan, Richard Caruana, Michael J Fine, Janine E Janosky, Gary Livingston, Tom Mitchell, et al. Predicting dire outcomes of patients with community acquired pneumonia. *Journal of biomedical informatics*, 38 (5):347–366, 2005.

Peng Ding and Fan Li. Causal inference: A missing data perspective. *Statistical Science*, 33(2):214–237, 2018.

Ofer Harel and Xiao-Hua Zhou. Multiple imputation: review of theory, implementation and software. *Statistics in medicine*, 26(16):3057–3077, 2007.

Trevor Hastie and Robert Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.

Joseph G Ibrahim, Haitao Chu, and Ming-Hui Chen. Missing data in clinical studies: issues and methods. *Journal of clinical oncology*, 30(26):3297, 2012.

Helena Jeličić, Erin Phelps, and Richard M Lerner. Use of missing data methods in longitudinal studies: the persistence of bad practices in developmental psychology. *Developmental psychology*, 45 (4):1195, 2009.

David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. *Logistic regression*. Springer, 2002.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR, 2017.

Sanjay Krishnan and Eugene Wu. Alphaclean: Automatic generation of data cleaning pipelines. *arXiv preprint arXiv:1904.11827*, 2019.

Jiang Li, Xiaowei S Yan, Durgesh Chaudhary, Venkatesh Avula, Satish Mudiganti, Hannah Husby, Shima Shahjouei, Ardavan Afshar, Walter F Stewart, Mohammed Yeasin, et al. Imputation of missing values for electronic health record laboratory data. *NPJ digital medicine*, 4(1):147, 2021.

Peng Li, Elizabeth A Stuart, and David B Allison. Multiple imputation: a flexible tool for handling missing data. *Jama*, 314(18):1966–1967, 2015.

Jimmy Lin, Chudi Zhong, Diane Hu, Cynthia Rudin, and Margo Seltzer. Generalized and scalable optimal sparse decision trees. In *International Conference on Machine Learning*, pages 6150–6160. PMLR, 2020.

Wei-Chao Lin and Chih-Fong Tsai. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53(2): 1487–1509, 2020.

Roderick JA Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, 83 (404):1198–1202, 1988.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158, 2012.

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–631, 2013.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322, 2010.

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019. Accessed at https://interpret.ml.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Mohammed Saeed, Christine Lieu, Greg Raber, and Roger G Mark. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. In *Computers in cardiology*, pages 641–644. IEEE, 2002. Accessed at https://archive.physionet.org/mimic2/ .

Yulia Sidi and Ofer Harel. The treatment of incomplete data: reporting, analysis, reproducibility, and replicability. *Social Science & Medicine*, 209:169–173, 2018.

Christian Steinruecken, Emma Smith, David Janz, James Lloyd, and Zoubin Ghahramani. The automatic statistician. In *Automated Machine Learning*, pages 161–173. Springer, Cham, 2019.

Daniel J. Stekhoven and Peter Bühlmann. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 10 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr597. URL https://doi.org/10.1093/bioinformatics/btr597.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

United States Department of Health and Human Services (US DHHS), Centers of Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS), and Division of Vital Statistics (DVS). *Birth Cohort Linked Birth – Infant Death Data Files, 2004-2015*, 2021. Compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program, on CDC WONDER On-line Database. Accessed at https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm.

Stef Van Buuren. *Flexible imputation of missing data*. CRC press, 2018.

Stef Van Buuren and Karin Groothuis-Oudshoorn. MICE: Multivariate imputation by chained equations in R. *Journal of statistical software*, 45:1–67, 2011.

Akbar K Waljee, Ashin Mukherjee, Amit G Singal, Yiwei Zhang, Jeffrey Warren, Ulysses Balis, Jorge Marrero, Ji Zhu, and Peter DR Higgins. Comparison of imputation methods for missing laboratory data in medicine. *BMJ open*, 3(8):e002847, 2013.

Zijie J Wang, Alex Kale, Harsha Nori, Peter Stella, Mark Nunnally, Duen Horng Chau, Mihaela Vorvoreanu, Jennifer Wortman Vaughan, and Rich Caruana. GAM changer: Editing generalized additive models with interactive visualization. *arXiv preprint arXiv:2112.03245*, 2021.

Jinsung Yoon, James Jordon, and Mihaela Schaar. GAIN: Missing data imputation using generative adversarial nets. In *International Conference on Machine Learning*, pages 5689–5698. PMLR, 2018.

Hsiang-Fu Yu, Nikhil Rao, and Inderjit S Dhillon. Temporal regularized matrix factorization for high-dimensional time series prediction. *Advances in neural information processing systems*, 29, 2016.

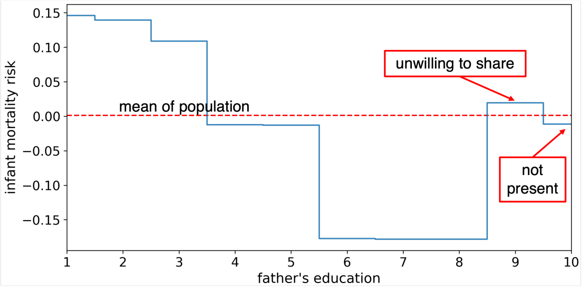


Figure 7: Impact of father's education on infant mortality risk, 2013.

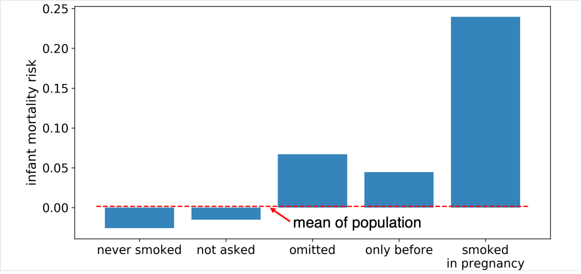


Figure 8: Impact of smoking before and during pregnancy on infant mortality risk, 2013.

Appendix A. Testing for MCAR with EBM: Case Study

In some cases we have information about the mechanism generating missing values and the likelihood that a similar mechanism will generate data in the future.

As an example, consider CDC Birth Cohort Linked Birth – Infant Death Data Files United States Department of Health and Human Services (US DHHS) et al. (2021). The dataset describes pregnancy and birth variables for all live births in the U.S. together with an indication of an infant's death before the first birthday. The dataset is collected using two certificates: 1989 Revision of the U.S. Standard Certificate of Live Birth (unrevised) and the 2003 revision of the U.S. Standard Certificate of Live Birth (revised). As a result of the delayed, phased transition to the 2003 Certificate, the cohorts from 2004 to 2015 include data for reporting areas that use the newer 2003 revision along with data for reporting areas that still use the older 1989 Certificate (unrevised), with later years having a larger fraction of data corresponding to the 2003 revision. Values for variables that are present only in the 2003 certificate will be missing for areas using the earlier, 1989 certificate. In 2013, 10% of records come from such areas, the fraction is declining year to year and we can expect it to be even smaller in subsequent years.

Figure 7 shows the impact of father's education on infant mortality risk according to an EBM model trained on 2013 data. Values from 1 to 8 correspond to different levels of educational attainment, with 1 indicating 8th grade or less and 8 a doctorate or professional degree. The risk is high for levels 1-3, drops to just below the average risk for levels 3-4 (some college and associate degree) and even further for BA/BS, MA/MS and doctorate (levels 6-8)[2]. Level 9 indicates unwillingness to share this information and 10 corresponds to 10% of records where this variable was not present (version 1989). Level 9 is associated with slightly elevated risk; we may guess that fathers unwilling to share are more likely to be lower on the education scale. Level 10 is associated with risk slightly below average, which is surprising at first glance. Unlike for Level 9, the mechanism according to which the information is withheld is independent of the value of the variable in question (namely, the geographical area using an older version of the certificate). However, if the populations using the two certificate versions were coming from the same distribution, we would expect average risk (0 on the shape function) for this group. The MCAR test from Section 4.1.1 indicates these groups are statistically different from each other, suggesting social, demographic or other differences between these populations.

A similar picture emerges when we look at infant mortality as a function of mother smoking before and during pregnancy. The risk is highest for mothers who smoked during pregnancy, slightly elevated for those who smoked before pregnancy and lowest for mothers who never smoked. Risk for mothers who didn't share this information ('omitted') is clearly elevated. The group for whom the value is missing (older 1989 certificate, denoted 'not asked') has risk slightly lower than average (0). Again, risk different from average indicates a distribution shift with respect to the rest of the population, and we see that 'omitted' is different from 'not asked'.

If we were to train an infant mortality risk model on 2013 data and use it for prediction on data from subsequent years, we could run into the problem of

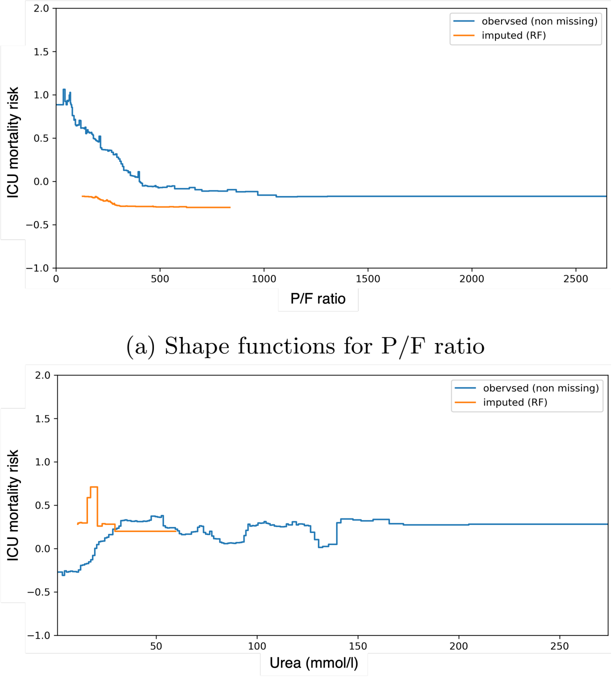
[2] Parents' education is the best proxy we have in the dataset for family's income.

values missing for an even lower fraction of all records and possibly coming from a distribution even more shifted with respect to the distribution of the majority of the records. Our model would likely predict the risk less accurately for this segment of the population.

Appendix B. Visualizing the Effect of Imputation

As mentioned in Section 5, the advanced imputation methods can significantly change the learned shape functions and such changes can sometimes be problematic. To help visualize the effect of imputation and identify potential problems in advance, we propose to separate the components of the missing group and the observed group in the EBM shape functions. To separate these two components, instead of directly imputing the missing values with the output of the imputation algorithm, we add a large offset to these imputed values so that the imputed values do not have overlap with the observed values. For example, in our experiments, we add max feature value plus 1 to the imputed values. This can be viewed as a trick to squeeze the feature and its missingness indicator variable into one dimension. Training EBMs on such separated feature values, the shape function will be a concatenation of the two curves corresponding to the observed group and the missing group. Also, because we know the offset we added to the imputed value, we can subtract it during visualization, and show the two curves on the same plot and original x-axis.

Figure 9 shows the EBM shape functions of the imputed group and the observed group separated using the method proposed above. Figure 9(a) shows that the risk of the RF imputed group is much lower than the risk of the observed group which corroborates what we found in Figure 6(a). Similarly, the effects of the imputed group in Figure 9(b) also differ significantly from the observed group, which explains why there exist spikes in the RF imputed EBM shape function in Figure 6(b). Using interpretable methods like EBMs allows one to understand the consequence of different imputation methods that otherwise would be invisible.



(a) Shape functions for P/F ratio

(b) Shape functions for Urea

Figure 9: EBM shape functions when the effects of imputation group (imputed by MissForest, denoted as RF imputed) and observed (non missing) groups are separated. The plots suggests how the two groups are different in terms of predicting the ICU mortality risk, and suggests how MissForest imputation might result in problematic models.