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

Alexander Hüdepohl

December 19, 2024

Outline

Motivation

Introduction

- missing values are a fundamental problem in data science and omnipresent in most datasets
- handling missing values can have a significant impact on ML models and corresponding results
- in medical applications, poor handling of missing data can lead to incorrect predictions and affect critical healthcare decisions

Challenges with missing values

- common preprocessing step: deleting rows or columns with missing values
- works only if:
  - missingness ratio is small
  - missing values are completely at random (MCAR)
- in other cases, deleting rows/columns may change the data distribution and introduce bias

Types of missingness

- missing values are categorized into three types:
  1. missing completely at Random (MCAR)
  2. missing at Random (MAR)
  3. missing not at Random (MNAR)
- different types of missingness require different handling methods:
  1. data cleaning and deletion
  2. imputation (e.g. mean, median, unique value)
  3. advanced methods like MICE, MissForest, and KNN Imputer

Why interpretability matters for missing values

- many imputation methods are black-box models
- users cannot easily recognize potential harms
- black-box models are hard to debug or explain
- interpretable ML provides new opportunities:
  - insights into missingness causes
  - detecting and avoiding risks
- glass-box models (e.g. EBM) combine high accuracy and interpretability

Related Work

Related work: Overview of methods

- generative imputation methods:
  - placing strong assumptions on underlying data distribution, not all are testable
  - can introduce bias
- discriminative imputation methods (MICE, MissForest, KNN Imputer)

Challenges in existing methods

- MissForest:
  - sensitive to initialized values
- KNN Imputer:
  - requires careful hyperparameter tuning (k)
  - struggles with high-dimensional datasets
- lack of interpretability:
  - cannot explain causes of missingness
  - no insights into imputation risks

Why transition to EBMs?

- provide a new perspective:
  - gain new insights on missingness mechanisms
  - better understand causes of missingness
  - detect risks introduced by imputation methods
- EBMs advantages:
  - visualize relationships via shape functions
  - identify anomalies such as spikes caused by imputation

Backround/Definitions

Definitions

- **missing completely at random (MCAR)**
  - the missingness is unrelated to the data (same for all samples)
  - example: A survey respondent accidentally skips a question
- **missing at random (MAR)**
  - the probability of missingness of a feature is determined from the observed values of the other features
  - example: Older patients are less likely to report their income
- **missing not at random (MNAR)**
  - the probability of missingness is also related to unobserved values in the data
  - example: Patients with depression avoid answering mental health questions

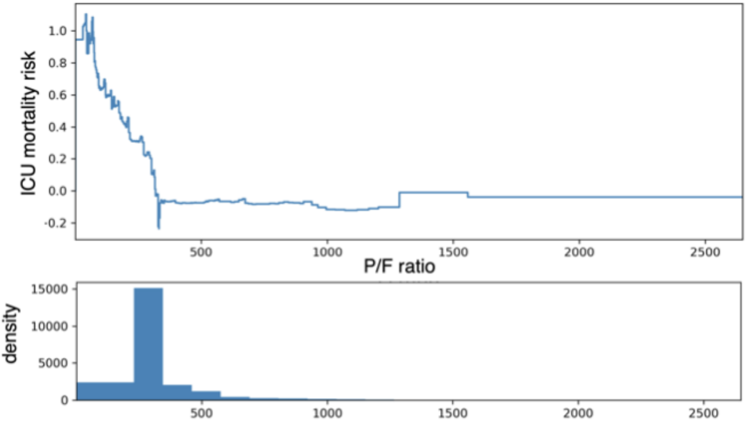
Missing value imputation

- **MissForest**
  - initial guess for missing value using mean imputation
  - sorts features according to missing rate
  - fits random forest iteratively to predict and impute each missing feature from other features until value converges
- **KNN Imputation**
  - imputes missing values by mean value of K nearest neighbors in training set
  - distance of two samples is measured on non-missing features on both samples
  - choose good distance metric and fine tune hyperparameter K

explainable boosting machines

- input sample denoted as (x,y)
  - x is p dimensional feature vector
  - y is the target
- **GAM:** $g(E[y]) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p)$
- β_0: Intercept
- $f_j's$: shape functions
- g: link function (identity for regression, logistic for classification)

EBM shape function and density plot



explainable boosting machines

- uses bagged ensembles of boosted depth-restricted tree to represent each f_j
- Tree-based ensemble learning improves performance of GAMs
- EBM's shape functions have more representational power and better capture fine detail (GAM uses splines)
- EBM improves accuracy by adding pairwise interactions: $g(E[y]) = \beta_0 + \sum_{j=1}^{p} f_j(x_j) + \sum_{k=1}^{K} f_k(x_{k1}, x_{k2})$

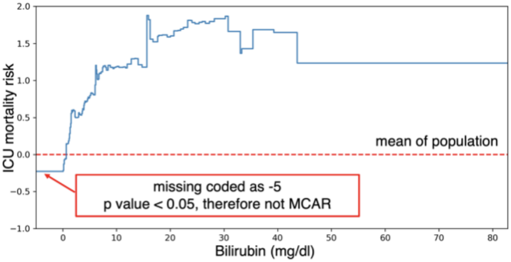
Testing for MCAR with EBM

Testing for MCAR with EBM

- important to determine type of missingness (e.g. Little-test for MCAR)
- encode missing values with a unique value for the feature
- fit EBM that predicts the target and get a shape function
- leaf nodes split feature values into many bins, each bin has a prediction score (bins and scores form shape function)
- shape function rewritten as linear combination of a series of indicator variables denoting if feature values are withing bin, coefficient are corresponding scores of the bins: $f_j(x_j) = \sum_{k=0}^{B_j-1} \theta_{j,k} \cdot \mathbb{1}\{b_{j,k} < x_j \leq b_{j,(k+1)}\}$,
- wald test (null hypothesis), shape functions centered with mean 0, if rejected we do not have MCAR (p-value < 0.05)

Example & Comparison

EBM shape function to test for MCAR



Comparison with Little's method

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

Missing Values in Healthcare Data

Missing values assumed normal

- in healthcare, it is common for feature values to be missing, because clinicians believed the patient was likely to be "normal"
- no patients in the data set with heart rates between 38 and 125; 91% are missing (coded as 0)
- when drawing the shape function, EBM will still make predictions in the range of missing values by interpolating from other bins

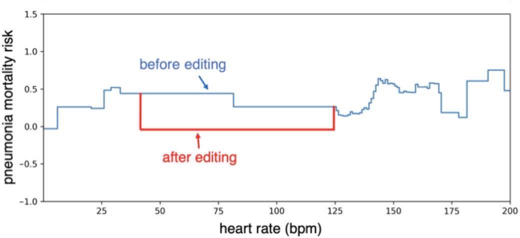
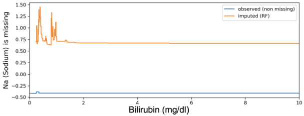
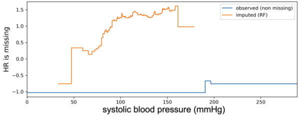


Figure: EBM shape function for heart rate (bpm)

Predicting the missingness (case MAR, MNAR)



(a) "Bilirubin" shape function when predicting missingness of "Na"



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

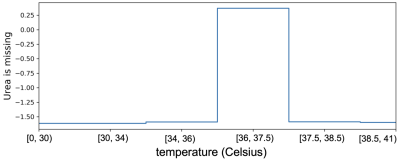


Figure: EBM shape functions for predicting the missingness of one feature using the others (x-axis: feature value, y-axis: contribution to missingness). The effects of the imputed group (orange) and the non-missing group (blue) are separated.

Test accuracy of predicting 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

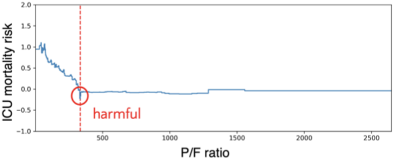
model	p_m	linear	curvilinear	quadratic
LR	0.1	0.957±0.013	0.901±0.013	**0.886±0.017**
RF		0.944±0.013	0.948±0.011	**0.886±0.017**
KNN		0.899±0.012	0.898±0.01	0.885±0.018
EBM		**0.959±0.012**	**0.963±0.011**	0.885±0.017
LR	0.2	0.928±0.018	0.847±0.035	0.817±0.010
RF		0.910±0.016	0.933±0.016	**0.828±0.012**
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

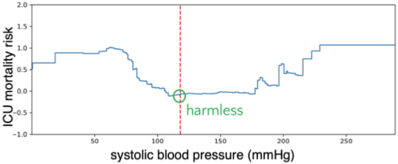
(b) datasets generated by MNAR

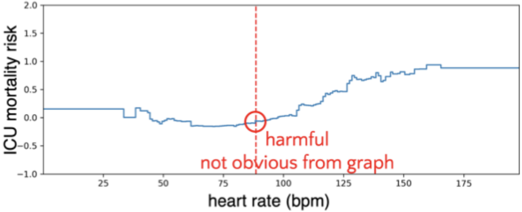
Detecting and avoiding potential risks of missing value imputations

Problem with bad imputations

- false imputations can influence references in a harmful way
  - for example creating artificial spikes in shape functions







Identifying harmful imputations

- 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
- first problem is easy to solve:
  - mean value is the same before and after mean imputation
  - Find the bin which covers the mean value and detect if that bin is a spike

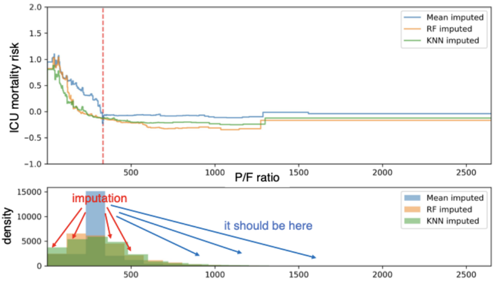
Procedure for detecting harmful imputations

- regarding the second problem:
  - need an algorithm to distinguish between spikes from mean imputation and natural fluctuations in shape functions
  - use second-order differences for all bins (spikes have extreme second-order differences)
  - run an outlier detection algorithm on the second-order differences (isolation forest)
- the algorithm predicts an anomaly score for each bin, and we choose a threshold so that around 5% of bins are detected as outliers
- potentially harmful mean imputations are predicted if bins covering the mean values are also detected as outliers

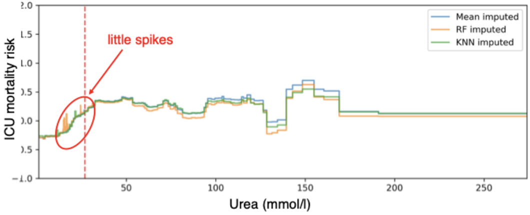
Challenges with advanced imputation methods

- MissForest struggles with systematic biases:
    - missing P/F ratio values for healthy patients lead to no real "healthy" observations in the data
    - thus MissForest might incorrectly assume that missing P/F values are similar to observed low values, as it is trained only on observed data and cannot infer that missing values indicate health
- impact of advanced methods:
    - harmful imputations can reduce predicted risk for riskier low-P/F-ratio patients, potentially leading to inadequate care
    - fluctuations in imputations can cause "little spikes" in EBM shape functions

Examples with advanced method imputations



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



(b) Shape functions for Urea

Conclusion

Capabilities of EBMs in addressing missing values

- detect and understand causes of missingness
- detect and avoid potential risks introduced by imputation methods:
  - predict missingness of features from other features
  - EBM interpretability helps users better understand the relationship between features and missingness
- editable to fix issues with minimal accuracy impact:
  - edits only affect small subsets of the model, preserving overall accuracy

Questions?

Thank you for your attention!
Any questions?