Missing Values Handling for Machine Learning Portfolios

Andrew Y. Chen, Jack McCoy

Journal of Financial Economics, 2024

Interpreter: Weihao Zhao

School of Economics and Management WHU

2024-10-16

Background

- Applying Machine Learning method to Asset Pricing.
- Combining many cross-sectional predictors(Economic gains).
- Missing at Random v.s. Missing Not at Random

Introduction – Motivation

1. **Missing values are a pervasive issue in financial datasets.**
   - Rare studies focus on missing structure.
2. **Certain missing structure implies certain method:**
   - Complex imputation introduces noise;
   - Simple mean imputation is good enough.
3. **Identifying missing structures is crucial**
   - Missing Not at Random(Against Rubin 1976)

Introduction– Research Question

- **How to handle missing values in cross-sectional predictors for machine learning portfolios?**

Why simple mean imputaion is as good as EM imputation?

Introduction– Contribution

1. Literature on handling cross-sectional missing values

Prior: Different imputation methods, not modeling the missing structure.(Freyberger et al, 2023; Bryzgalova et al, 2023)

Extend: Compare different methods and explain missing structure.

2. Literature on using different dataset

Prior: Differ in predictor and stock return datasets.

Extend: Using CZ documentation;(Chen, Zimmerman, 2022)

The empirical results complement existing research.

Introduction– Contribution

3. Literature on using mean imputation in machine learning

Prior: Using complex mean imputation methods and many predictors.

Extend: Simple mean imputation method is helpful for transparency.

4. Literature on dimensionality of predictions

Prior: Debate on whether few factors explain cross-sectional returns.

Extend: Low correlation between predictors implies high dimensions.

Hypothesis

1. Simple mean imputation is effective enough.

The missingness structure(CZ, 2022).

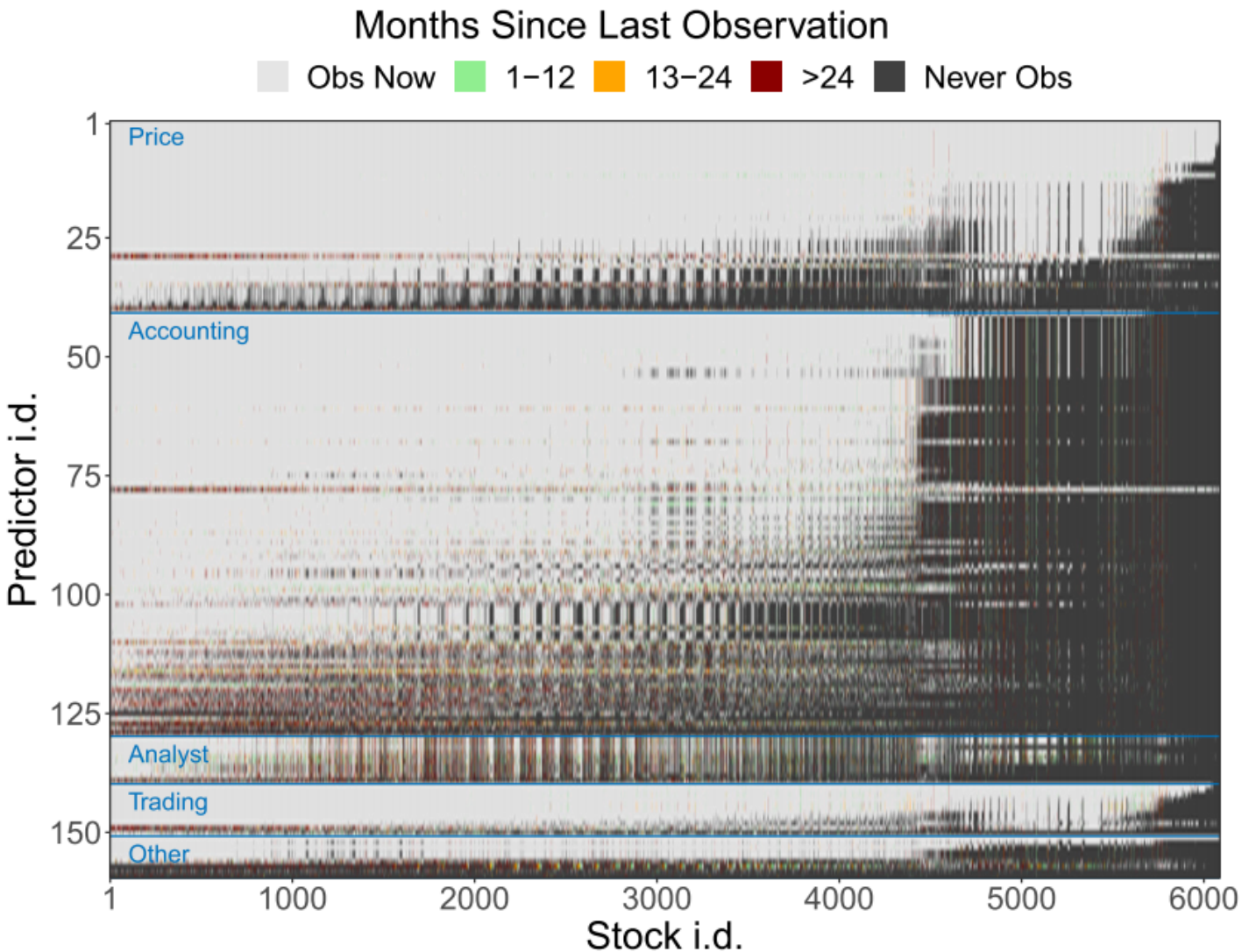
2. EM imputation underperforms under certain settings.

The covariance structure.

Missingness Structure(CZ, 2022)

1. **Missingness occurs in blocks organized by time**
2. **Block structures within data categories**
3. **Missingness occurs in blocks organized by the underlying data.**

Missingness Structure



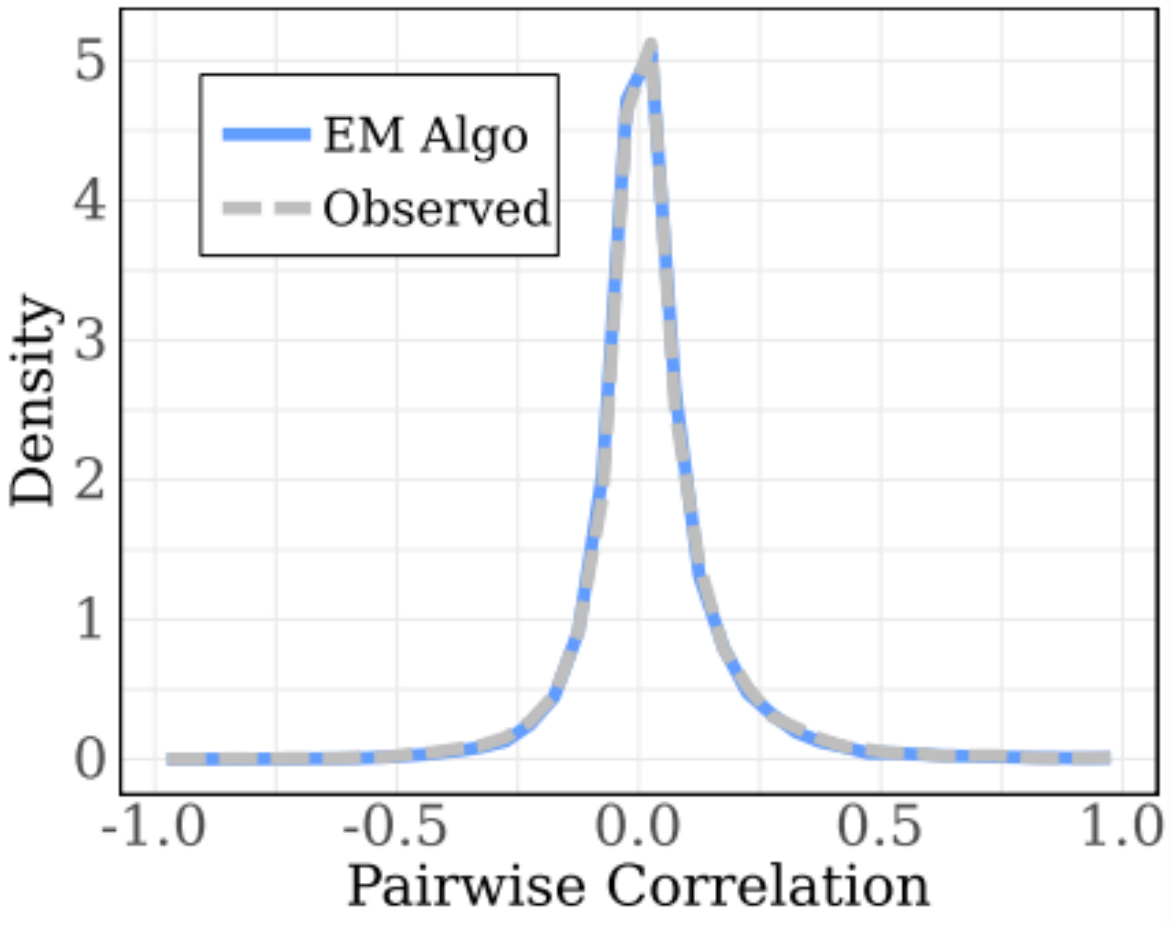
Sophisticated Imputation

EM imputation:

Formula: $\hat{X}_{\text{miss}|i,t} = \beta'_{i,t} X_{\text{obs}|i,t}$, where $\beta_{i,t} = \Sigma^{-1}_{\text{obs,obs}|i,t} \Sigma_{\text{obs,miss}|i,t}$.

- No significant performance improvement;
- Σ_t are mostly close to zero (CZ, 2022).

Distribution of Correlations for EM Algo



Corr between predictors, the density map shows a peak near 0.

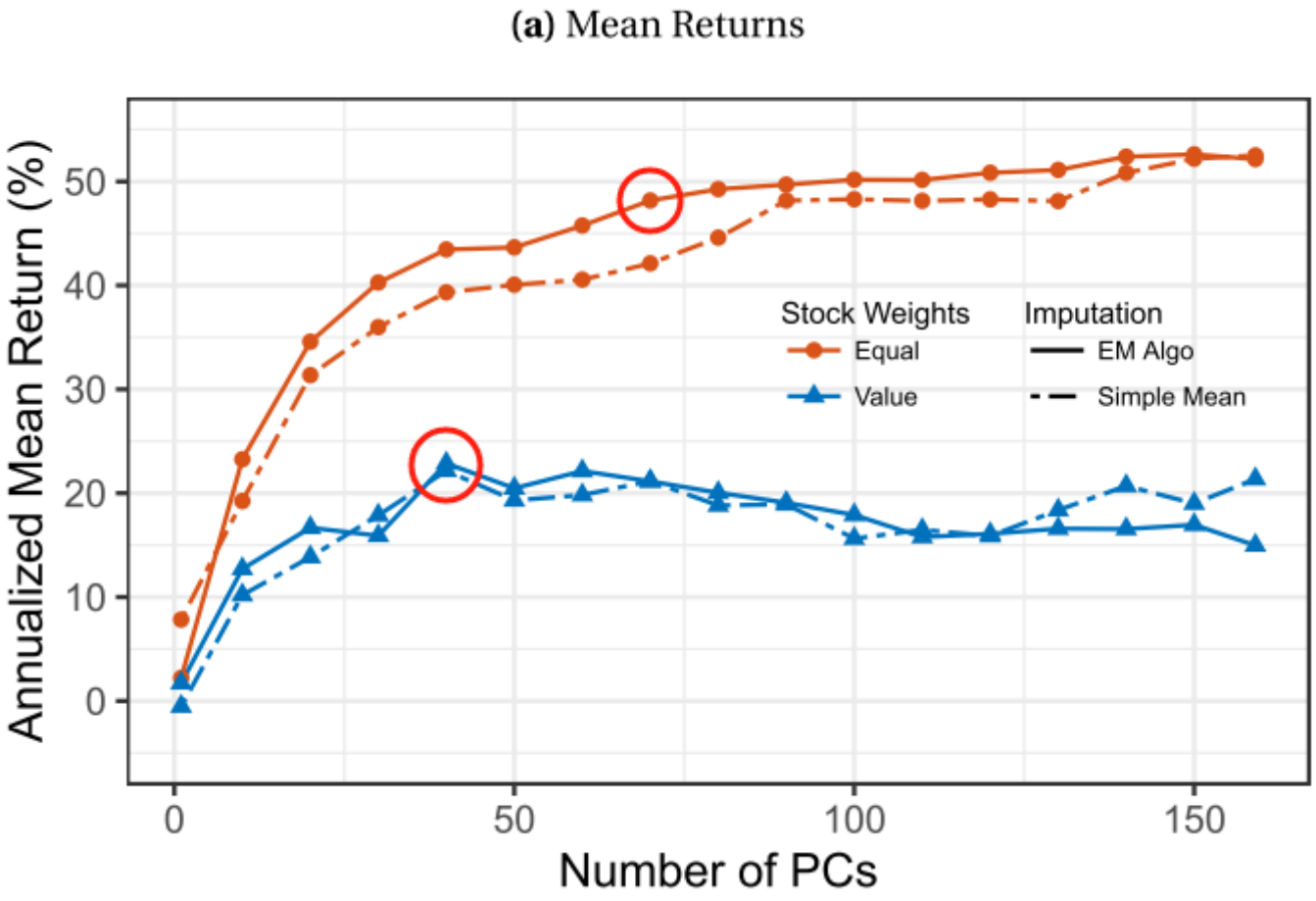
Research Design – Methodology

1. **Comparing imputation methods**:
   - Simple mean imputation vs. Complex imputation (e.g. EM).
2. **Machine learning models constructing portfolios**:
   - PCA, sPCA, (GBRT, Neural Networks for robustness).

Research Design – Data

1. 159 cross-sectional stock return predictors.
2. Data sourced from CRSP and Compustat, 1985–2021.
3. **CZ documentation as a reference** :
   - 212 predictors from 153 literature;
   - Showing reproducibility(open-source data and code);
   - Low correlation between most predictors.

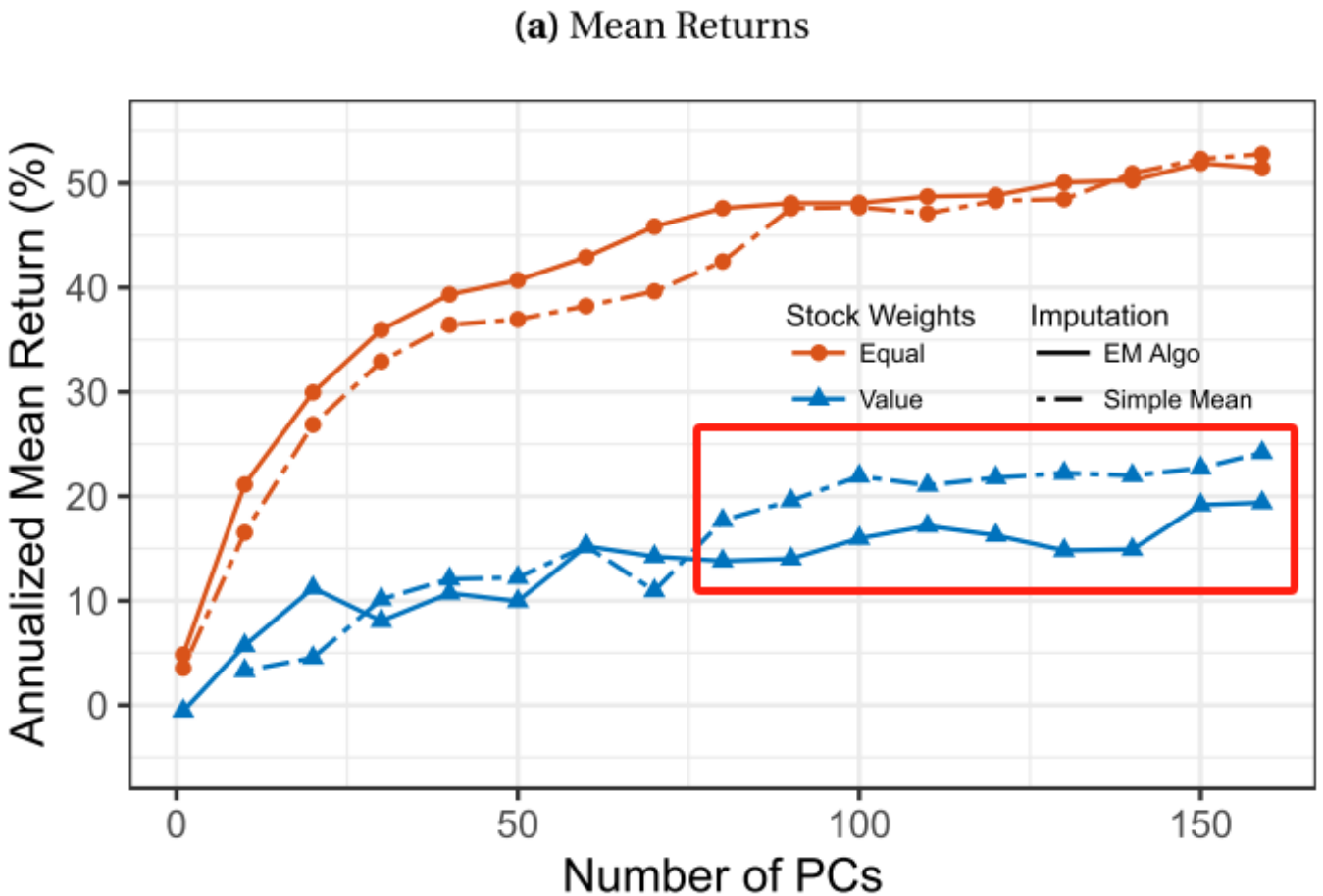
PCA Results – Simple v.s. Sophisticated(By Size)



- About 40 PCs for **vw** and 70 PCs for **ew**.

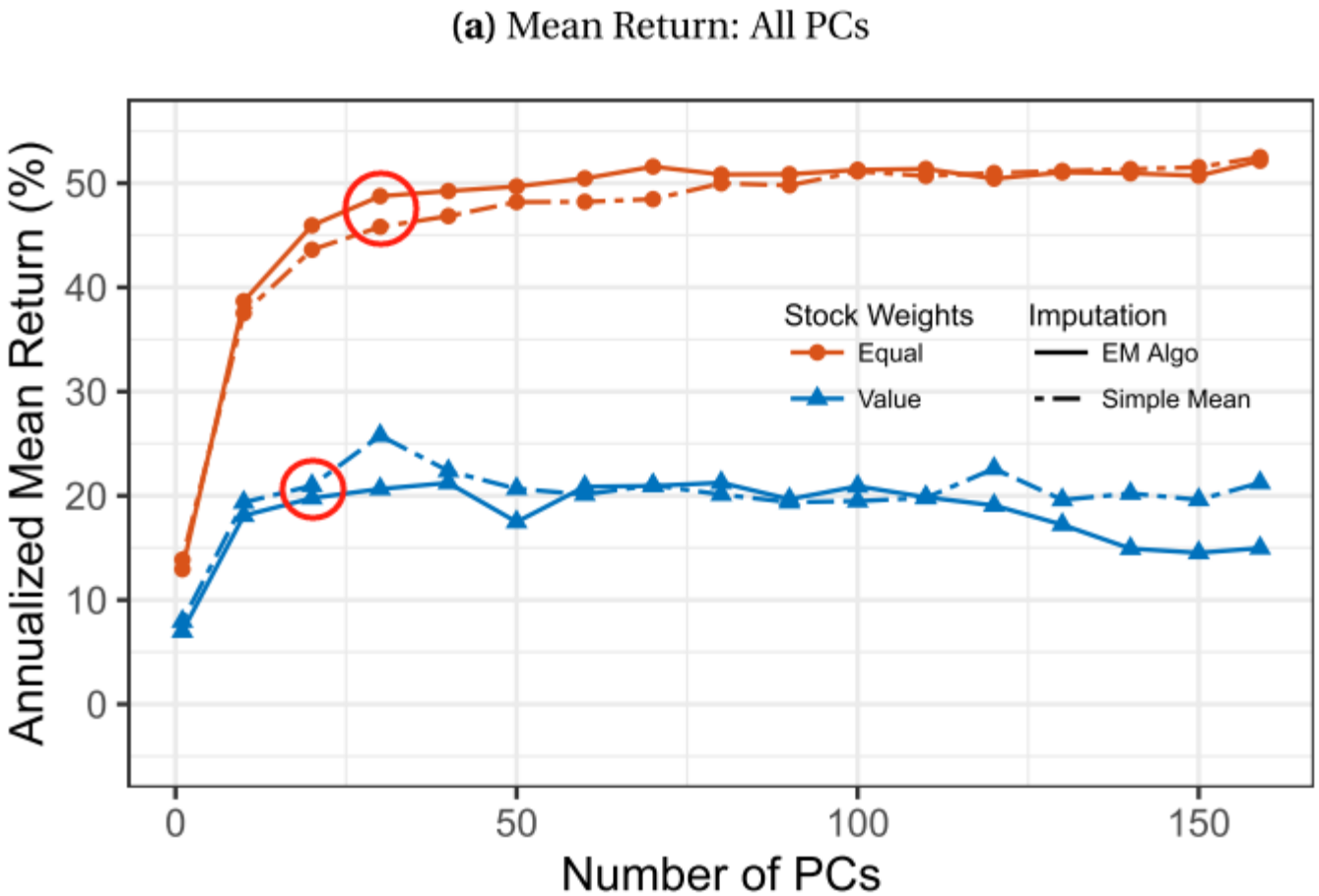
PCA Results – Simple v.s. Sophisticated(Pool)

- Poor performance in small cap stocks(More unstable cov structure)



sPCA for Dimensionality problem

- Reduces PCs unrelated to returns, which reducing dimensions.



Robustness Checks

- Mean imputation shows consistent performance across ML models.

Panel (b): Value-Weighted Mean Return (% Annualized)

	Forecasting Methods											
	Linear						*Non-Linear*					
	OLS		PCR		sPCR		GBRT		NN1		NN3	
	By Size	*Pool*	By Size	*Pool*	By Size	*Pool*	By Size	*Pool*	By Size	*Pool*	By Size	*Pool*
Ad-hoc Imputations												
Mean	31	*27*	32	*21*	27	*23*	17	*25*	40	*37*	37	*40*
Ind / Size	31	*27*	31	*21*	28	*22*	22	*12*	37	*37*	34	*40*
Last Obs	30	*26*	38	*20*	34	*21*	23	*7*	38	*38*	35	*45*
EM Imputations												
EM	28	*19*	36	*16*	31	*18*	32	*11*	43	*34*	39	*40*
EM AR1	34	*21*	35	*18*	33	*18*	27	*13*	39	*33*	36	*40*
pPCA10	30	*23*	34	*20*	31	*18*	32	*17*	40	*33*	38	*40*

Conclusion

1. **Finding missingness is important for handling missing data**
2. **Mean imputation: Effective enough, simple method.**
   - Less noise than advanced methods and is robust across models.
3. **The dimensionality problem exists if used many predictors.**
   - The use of sPCA reduces the need for common predictors(PCs).

Discussion

1. **Explore applicability of mean imputation to other datasets.**
   - Bond market, foreign exchange market, commodity market, etc.
   - Characteristics and data structures of different markets vary.
2. **Testing other ML methods to handle missing values.**
   - K-NN algothrm, Navie Bayes, Random Forest;
   - Deep learning method: Datawig for imputation.