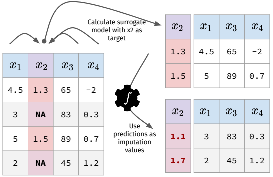
Applied Machine Learning

Imputation:
Advanced Methods and Pitfalls



Calculate surrogate model with x2 as target

x_1	x_2	x_3	x_4
4.5	1.3	65	-2
3	**NA**	83	0.3
5	1.5	89	0.7
2	**NA**	45	1.2

x_2	x_1	x_3	x_4
1.3	4.5	65	-2
1.5	5	89	0.7

Use predictions as imputation values

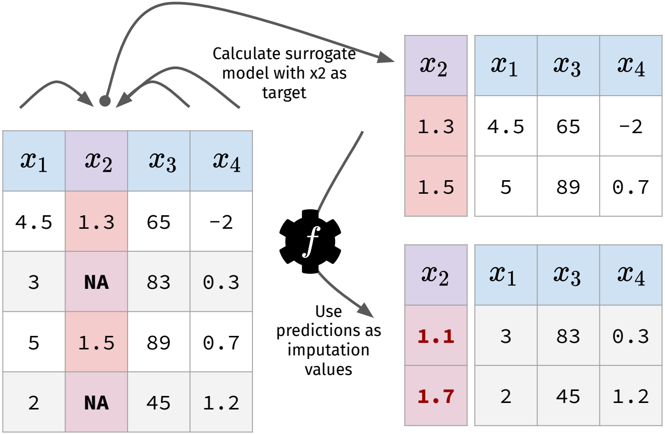
x_2	x_1	x_3	x_4
1.1	3	83	0.3
1.7	2	45	1.2

Learning goals

- Based imputation strategies
- Practical considerations for missing values

MODEL-BASED IMPUTATION

Instead of imputing a single value or sampling values it is desirable to take advantage of structure and correlation between features.



x_1	x_2	x_3	x_4
4.5	1.3	65	-2
3	**NA**	83	0.3
5	1.5	89	0.7
2	**NA**	45	1.2

Calculate surrogate model with x2 as target

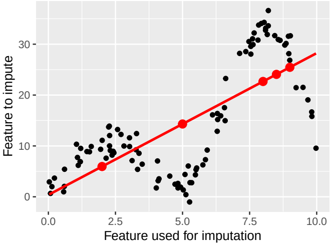
x_2	x_1	x_3	x_4
1.3	4.5	65	-2
1.5	5	89	0.7

Use predictions as imputation values

x_2	x_1	x_3	x_4
1.1	3	83	0.3
1.7	2	45	1.2

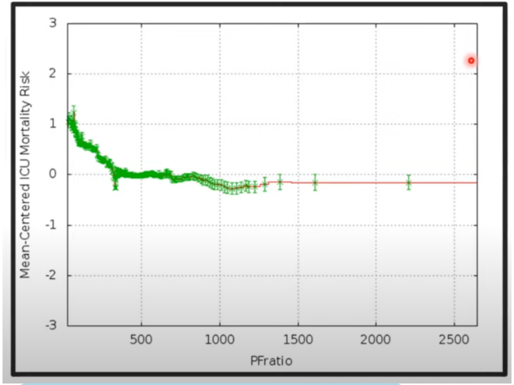
MODEL-BASED IMPUTATION: DRAWBACKS

- Choice of surrogate model has high influence on the imputation:



- Surrogate model should handle missing values itself, otherwise imputation *loop* may be necessary.
- Surrogate model hyperparameters can be tuned and can be different for each feature to impute.

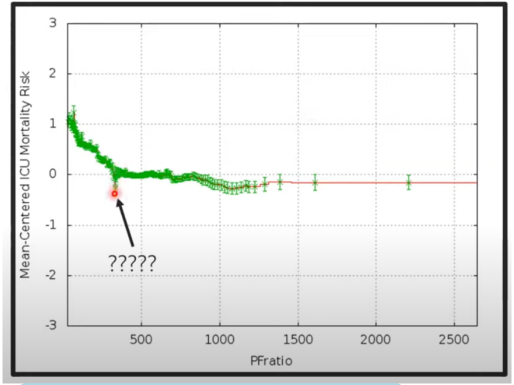
FAILED IMPUTATION: ICU OXYGENATION (I)



Source: ▸ *Rich Caruana's talk at AutoML Conference* 2024 (example at 43'31)

- ICU mortality vs. **PF ratio** (oxygenation)
- ≈ 0: no breathing; ~ 1000: healthy
- Expect *decreasing* risk with PF

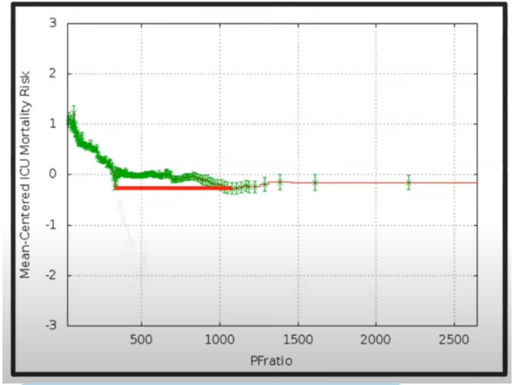
FAILED IMPUTATION: ICU OXYGENATION (II)



Source: ▸ *Rich Caruana's talk at AutoML Conference* 2024 (example at 43'31)

- Local dip in risk at one PF value
- Contradicts physiology and trend
- **Question:** Why so low?

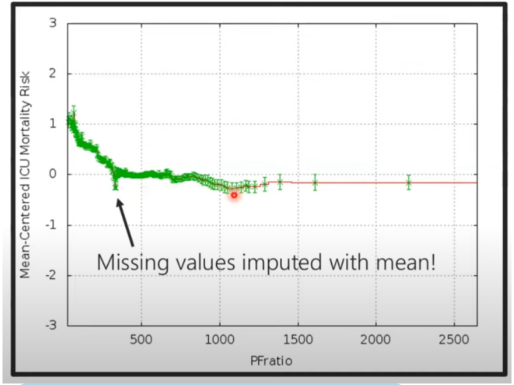
FAILED IMPUTATION: ICU OXYGENATION (III)



Source: ▸ *Rich Caruana's talk at AutoML Conference* 2024 (example at 43'31)

- Missing PF values imputed by the mean (collapse to **one** PF point)
- Model learns spuriously *low* risk there
- Imputation ignores MNAR/MAR structure

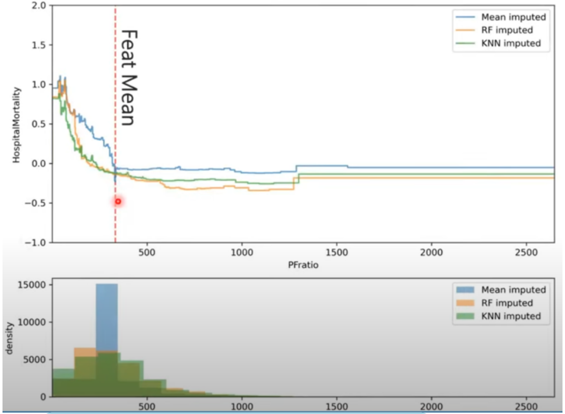
FAILED IMPUTATION: ICU OXYGENATION (IV)



Source: ▸ *Rich Caruana's talk at AutoML Conference* 2024 (example at 43'31)

- With imputation: $\sim 60\%$ healthy mapped to PF < 1000
- Label-feature mismatch near imputed PF
- Risk landscape shifts toward "healthier"
- **Should we use a more complex imputer?**

FAILED IMPUTATION: ICU OXYGENATION (V)



Source: ▸ *Rich Caruana's talk at AutoML Conference* 2024 (example at 43'31)

- "Smarter" imputers *spread* healthy into PF $<$ 1000
- Even worse: Final model underestimates risk for the sick
- Prefer **no fill** + missingness indicator
- If imputing: fit/tune *inside* CV; consider MNAR

TESTS FOR MISSINGNESS MECHANISMS

MAR vs **MCAR**

- Little’s test
- Logistic regression for missingness
  1. For each column, create a new classification task: predict if the value is missing or not
  2. Use the remaining columns as features
  3. Fit logistic regression
  4. Examine coefficients and p-values

MNAR: no test available

CENSORING MECHANISM

- **Left censoring:** True value below a threshold; exact value unknown.
- **Interval censoring:** True value lies within a known interval.
- **Right censoring:** True value above a threshold; exact value unknown.

CENSORING MECHANISM

- **Left censoring:** True value below a threshold; exact value unknown.
- **Interval censoring:** True value lies within a known interval.
- **Right censoring:** True value above a threshold; exact value unknown.
- **Type I censoring:** Experiment ends at pre-specified time; remaining subjects are right-censored.
- **Type II censoring:** Experiment ends after a fixed number of failures; remaining subjects are right-censored.
- **Random (non-informative) censoring:** Censoring times independent of failure times; observe minimum of failure and censoring time.

TAKEAWAYS OF A RECENT BENCHMARK

Imputation for prediction: beware of diminishing return. Le Morvan and Varoquaux, ICLR 2025 ▸ Morvan and Varoquaux 2024

- Better imputation performance does not result in better prediction performance for MNAR (w/o missingness indicator)
- Better imputation might be irrelevant b/c information is available in other features, unimportant, even with imputation, it might be hard to learn a good downstream model
- More expressive models benefit less from imputation
- Best on average: missForest + XGBoost + masking
- Open Questions:
  - performance of random draws
  - performance of multiple imputation
  - impact of performance on explainability and calibration
  - performance of models that can learn directly with missing value (advanced transformer architectures)

TAKEAWAYS

- There exist different reasons for missing data
- There exist different missingness mechanisms
- Different downstream tasks ask for different imputation strategies
- Different data modalities ask for different imputation strategies
- Understanding the data generating process helps deciding on the imputation strategy
- Common imputation strategies:
  - Constant value: mean, median
  - Imputation by sampling
  - Missingness indicator
  - Model-based

  → imputation selection is a CASH problem