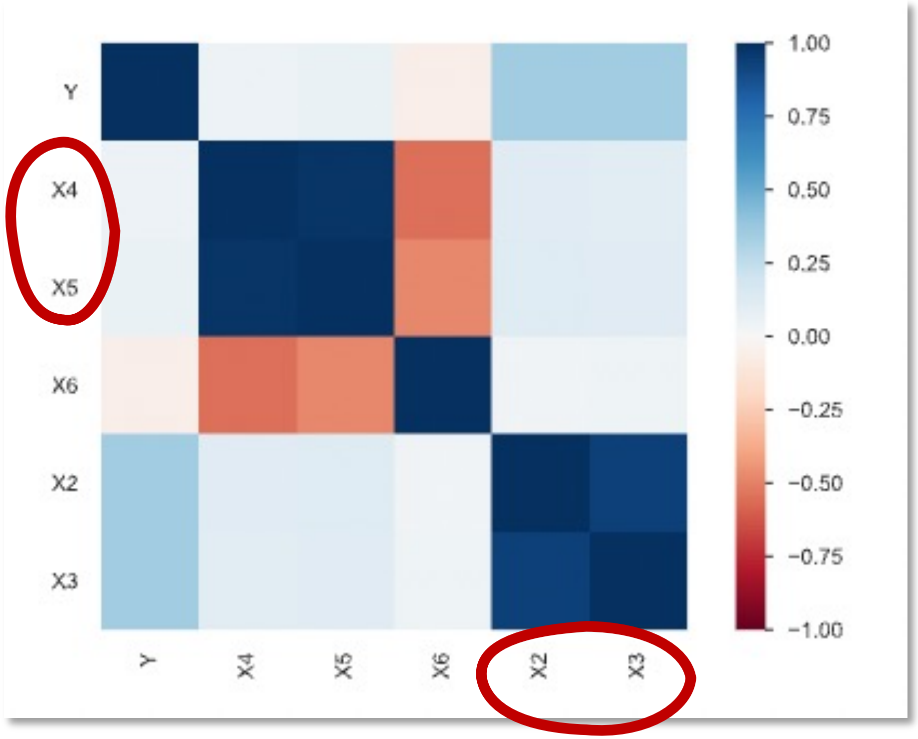
Biological Data CSV

Volker Hoffmann (volker@cheleb.net)

Data 1/2

Drop One?

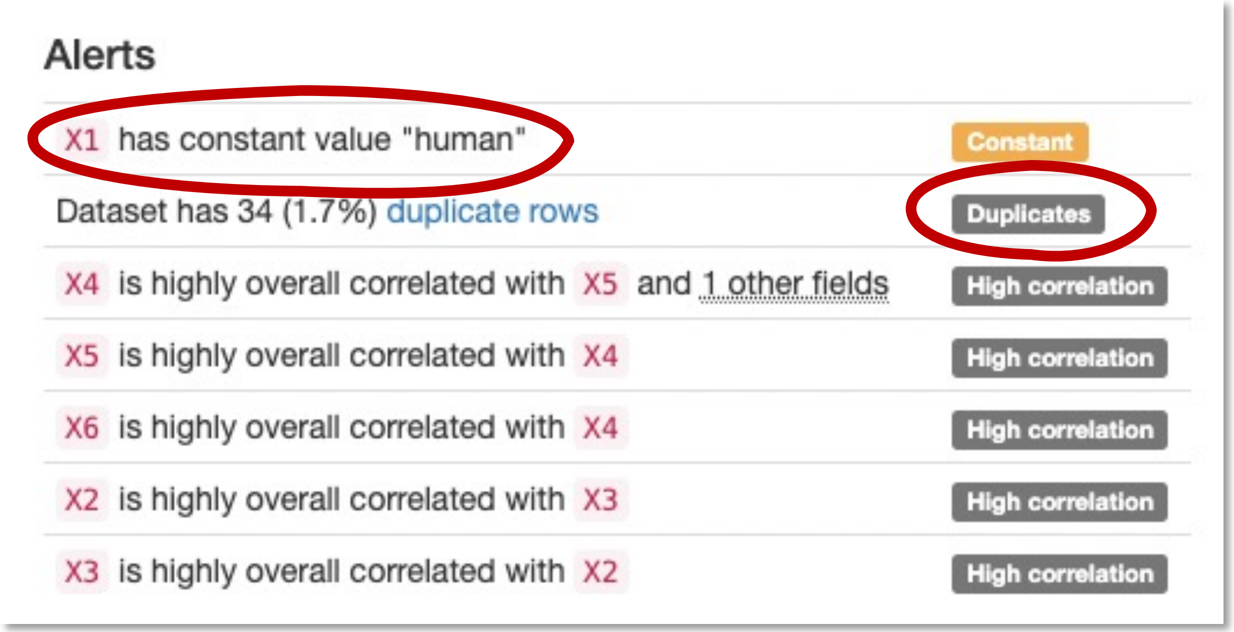
PCA?



Drop One?

Drop

Drop



Alerts

X1 has constant value "human" — Constant

Dataset has 34 (1.7%) duplicate rows — Duplicates

X4 is highly overall correlated with X5 and 1 other fields — High correlation

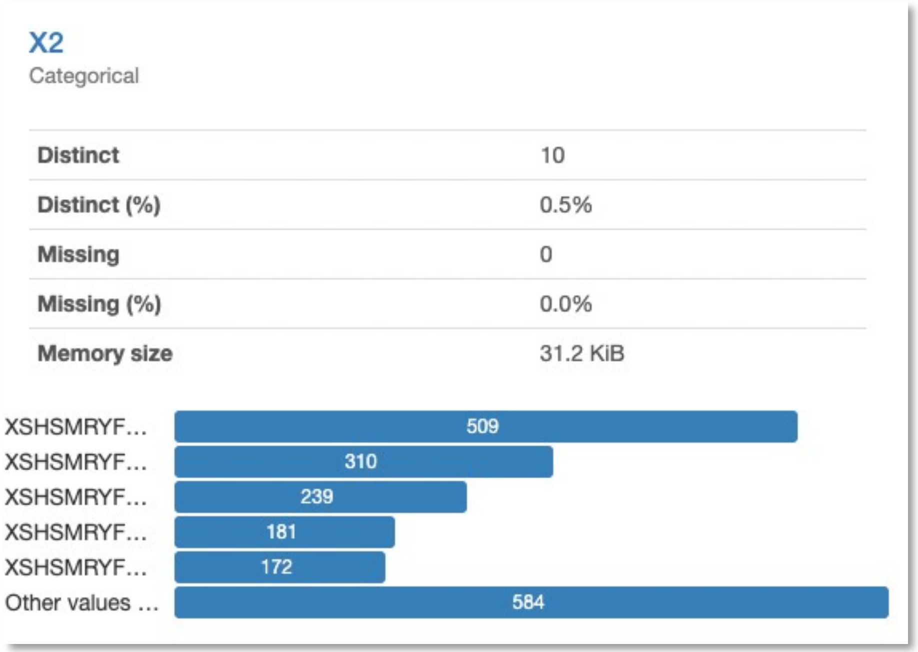
X5 is highly overall correlated with X4 — High correlation

X6 is highly overall correlated with X4 — High correlation

X2 is highly overall correlated with X3 — High correlation

X3 is highly overall correlated with X2 — High correlation

Data 2/2



X2

Categorical

Distinct	10
Distinct (%)	0.5%
Missing	0
Missing (%)	0.0%
Memory size	31.2 KiB

XSHSMRYF...	509
XSHSMRYF...	310
XSHSMRYF...	239
XSHSMRYF...	181
XSHSMRYF...	172
Other values ...	584

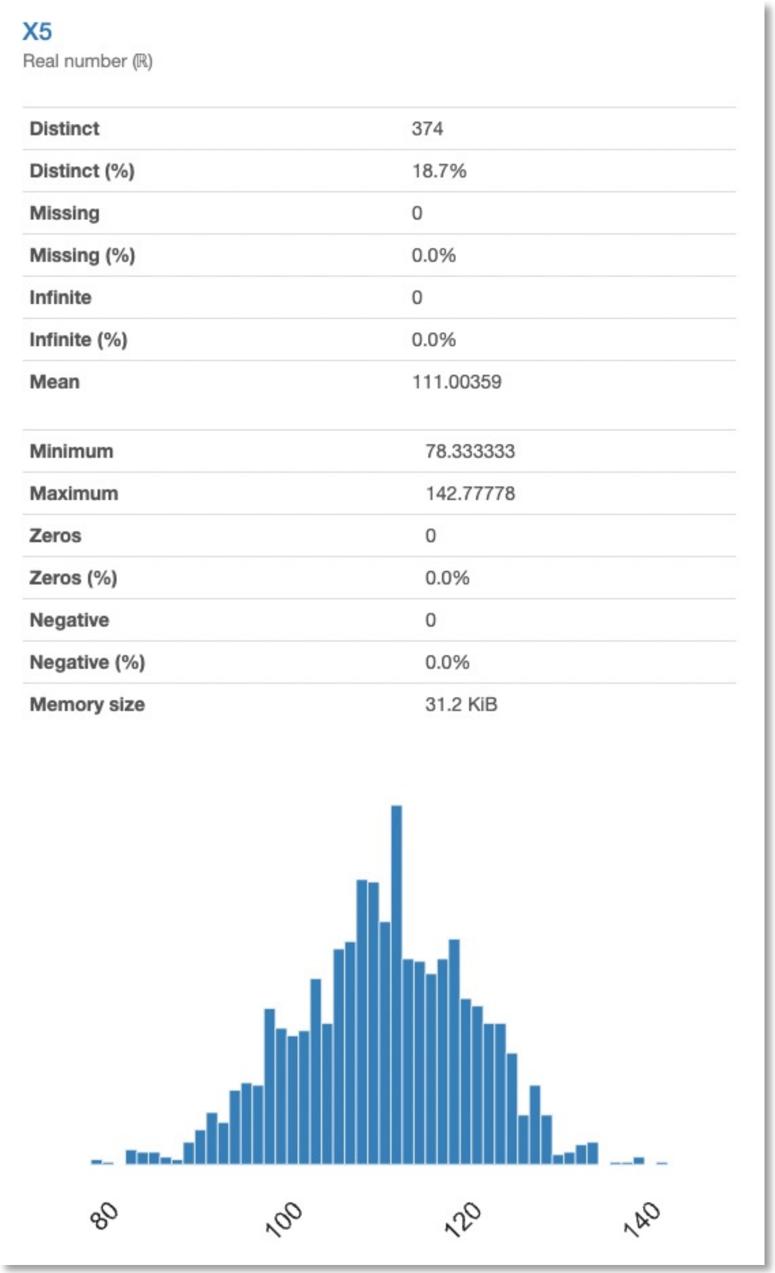
Encode as **One-Hot** / ~~Label~~

(Same for X3)

“Dummy Variable Trap”!

🦄 Np.rand.randn * 20 + 110 → ?

(X4, X6 also look like this)



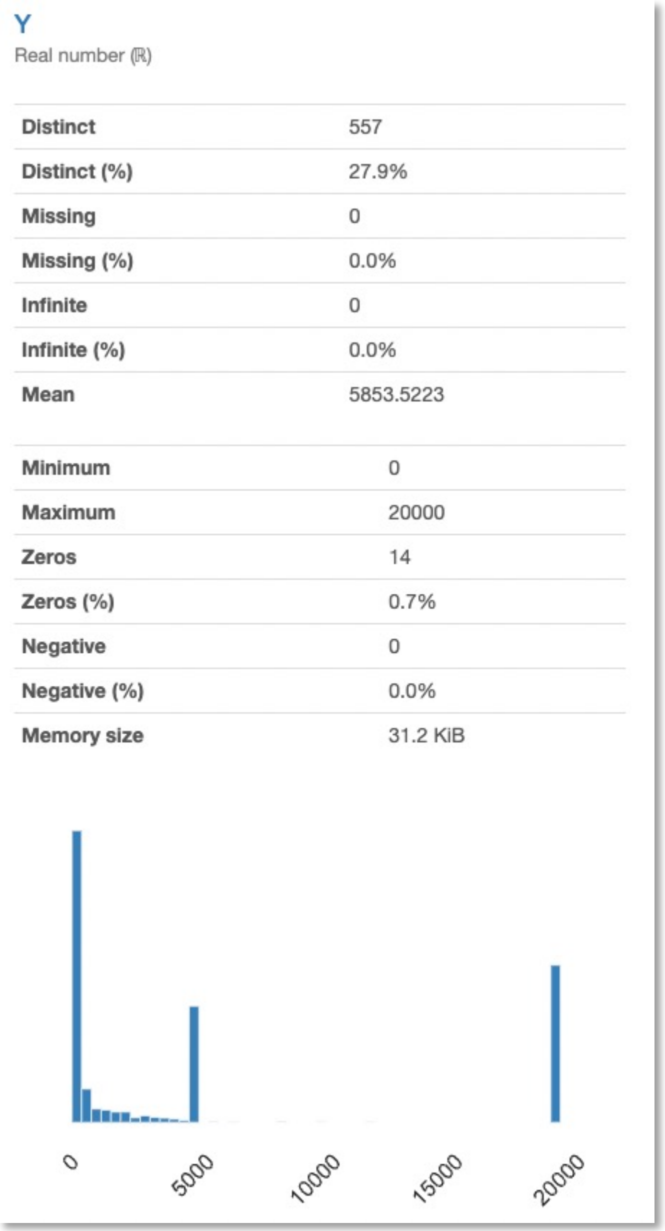
X5

Real number (ℝ)

Distinct	374
Distinct (%)	18.7%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	111.00359

Minimum	78.333333
Maximum	142.77778
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	31.2 KiB

Regression 1/3



Y

Real number (ℝ)

Distinct	557
Distinct (%)	27.9%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	5853.5223

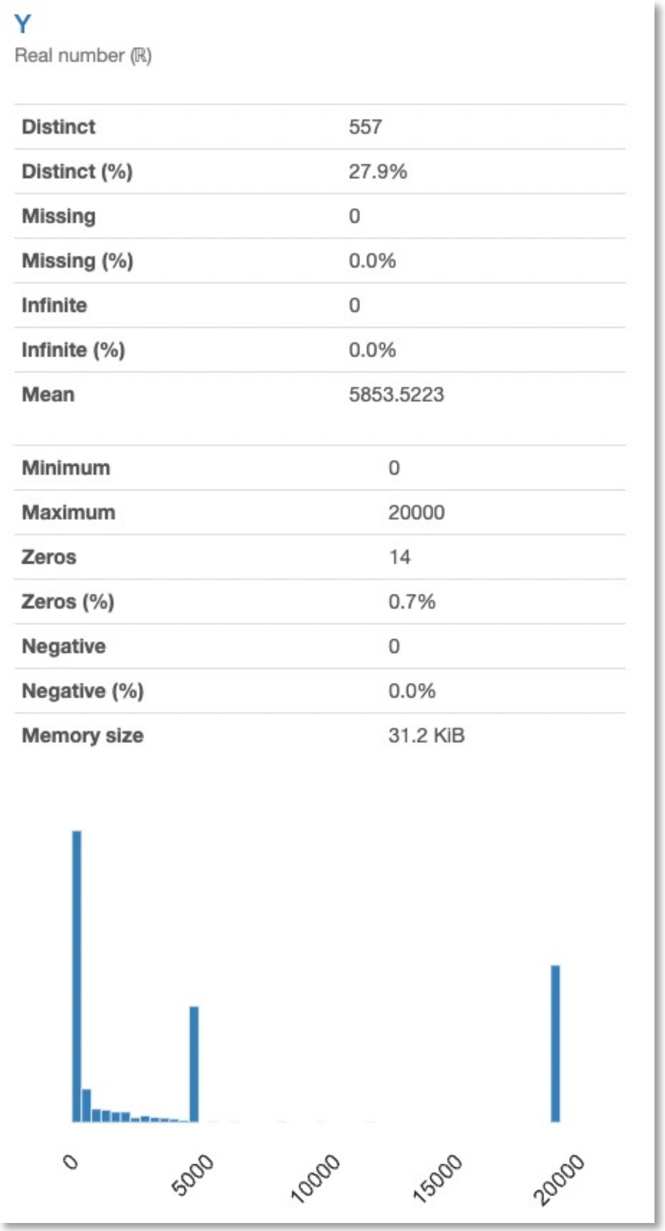
Minimum	0
Maximum	20000
Zeros	14
Zeros (%)	0.7%
Negative	0
Negative (%)	0.0%
Memory size	31.2 KiB

Regression 1/3

- Probably tricky… (look at histogram!)

- 3 Models
  - XGBoost
  - SVM
  - Ridge Regression (LinReg + Regularization)

- Rescaled Features to ~[0,1] (SVR/LinReg Care)
- 5-Fold Cross-Validation
- (Negative) Mean Squared Error



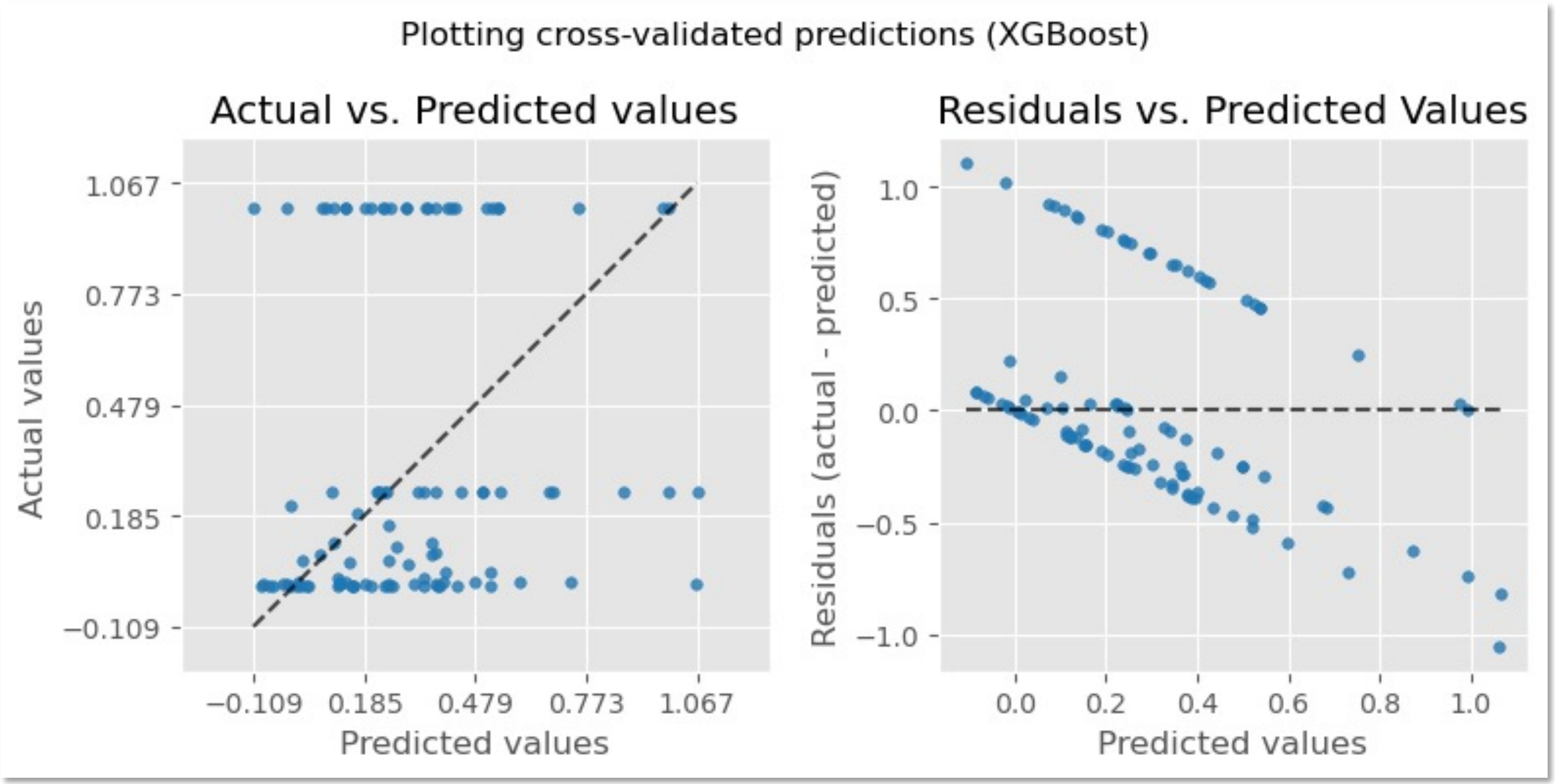
Y

Real number (ℝ)

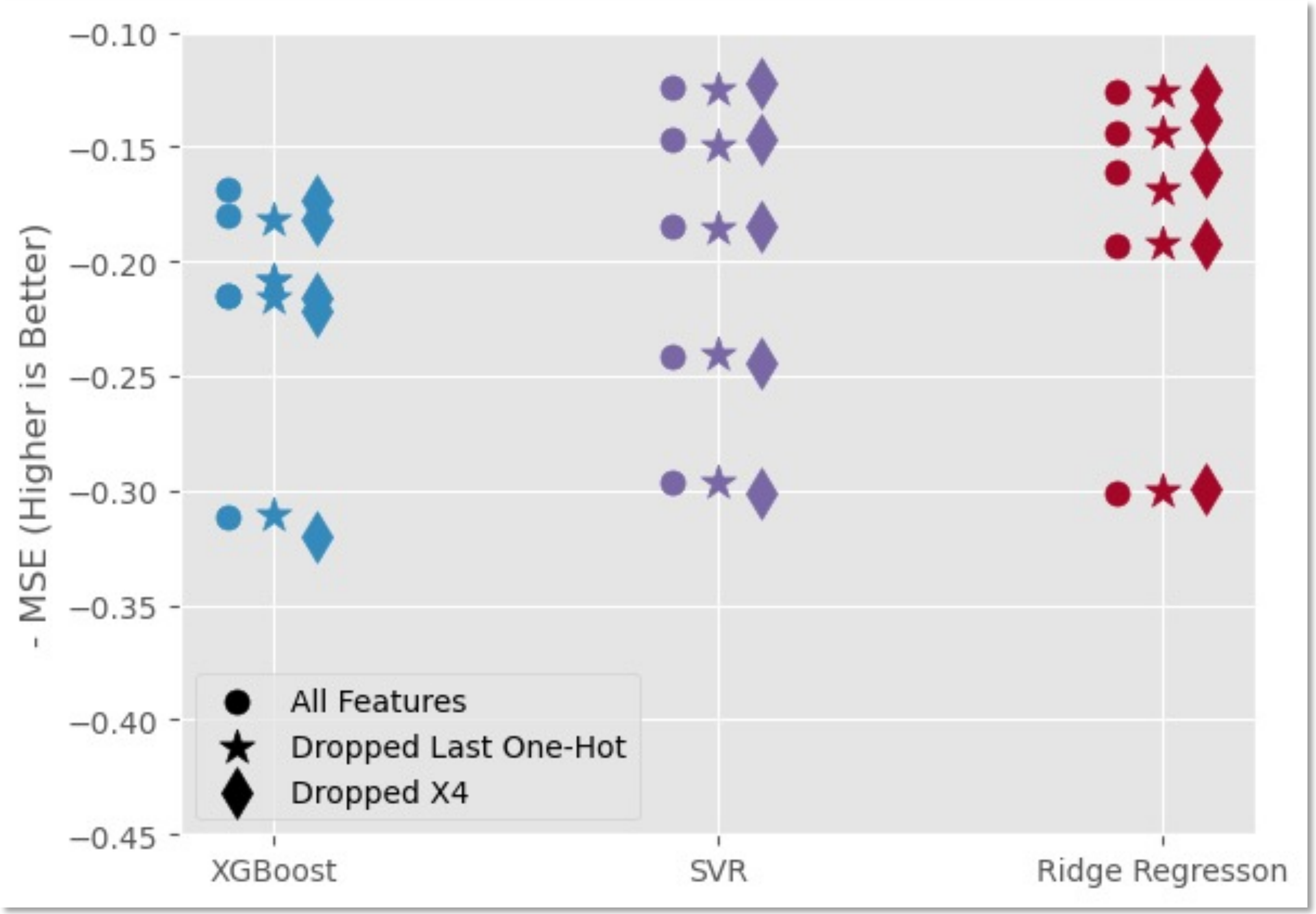
Distinct	557
Distinct (%)	27.9%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	5853.5223

Minimum	0
Maximum	20000
Zeros	14
Zeros (%)	0.7%
Negative	0
Negative (%)	0.0%
Memory size	31.2 KiB

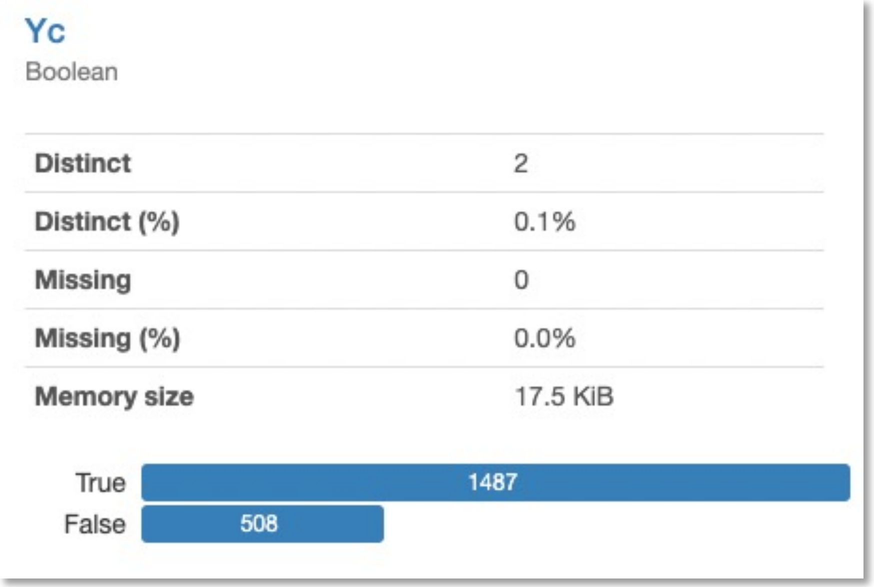
Regression 2/3



Regression 3/3



Classification 1/3



Yc

Boolean

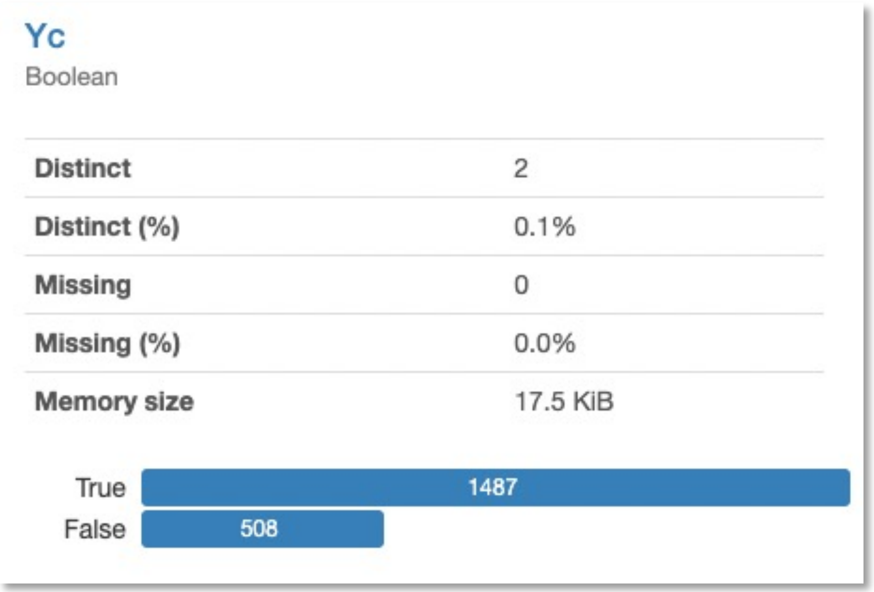
Distinct	2
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	17.5 KiB

Classification 1/3

- Probably easier
- Slightly imbalanced
  - Precision-Recall (instead of ROC Curve)

- 3 Models
  - XGBoost
  - SVC
  - LogReg

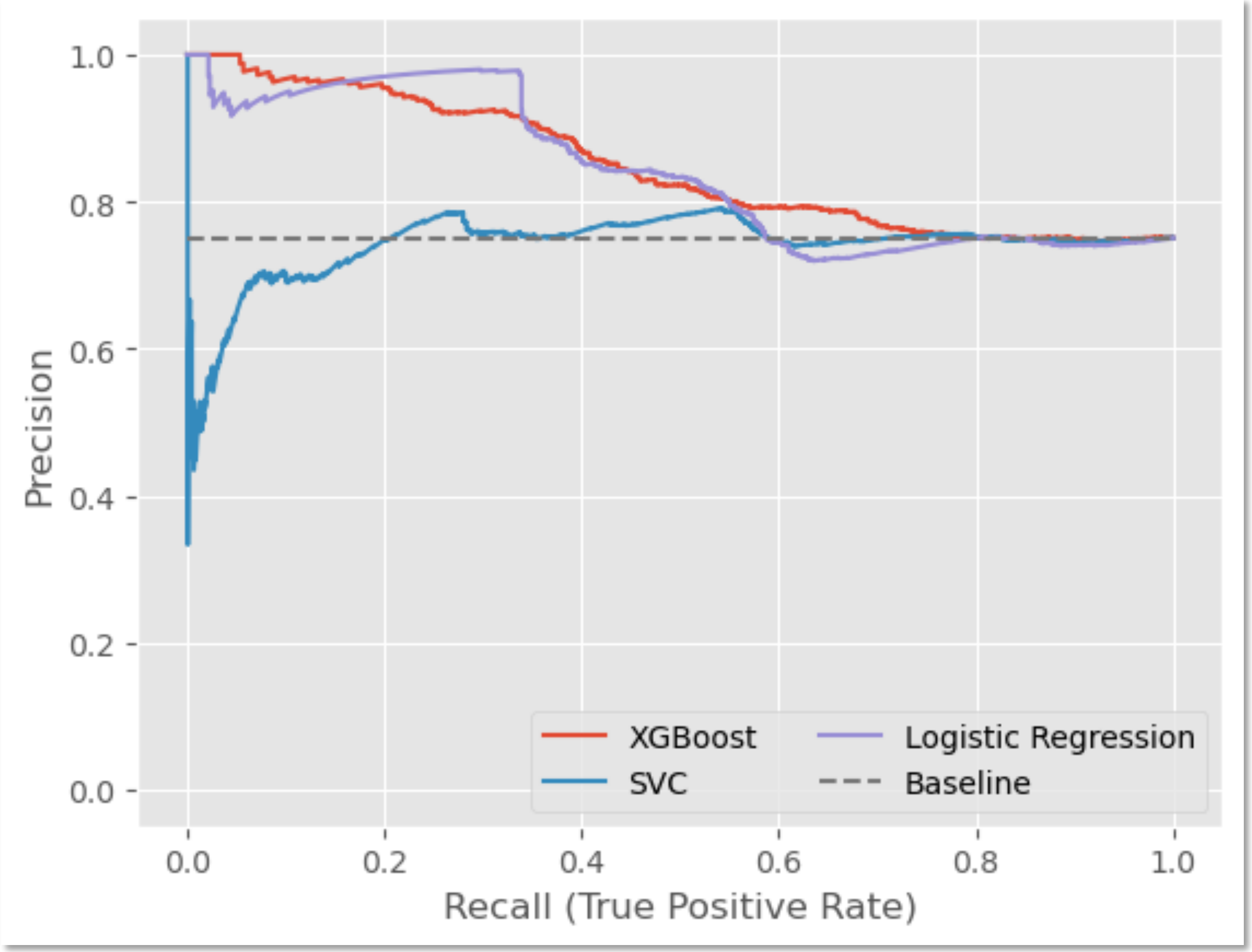
- 5-Fold Stratified CV



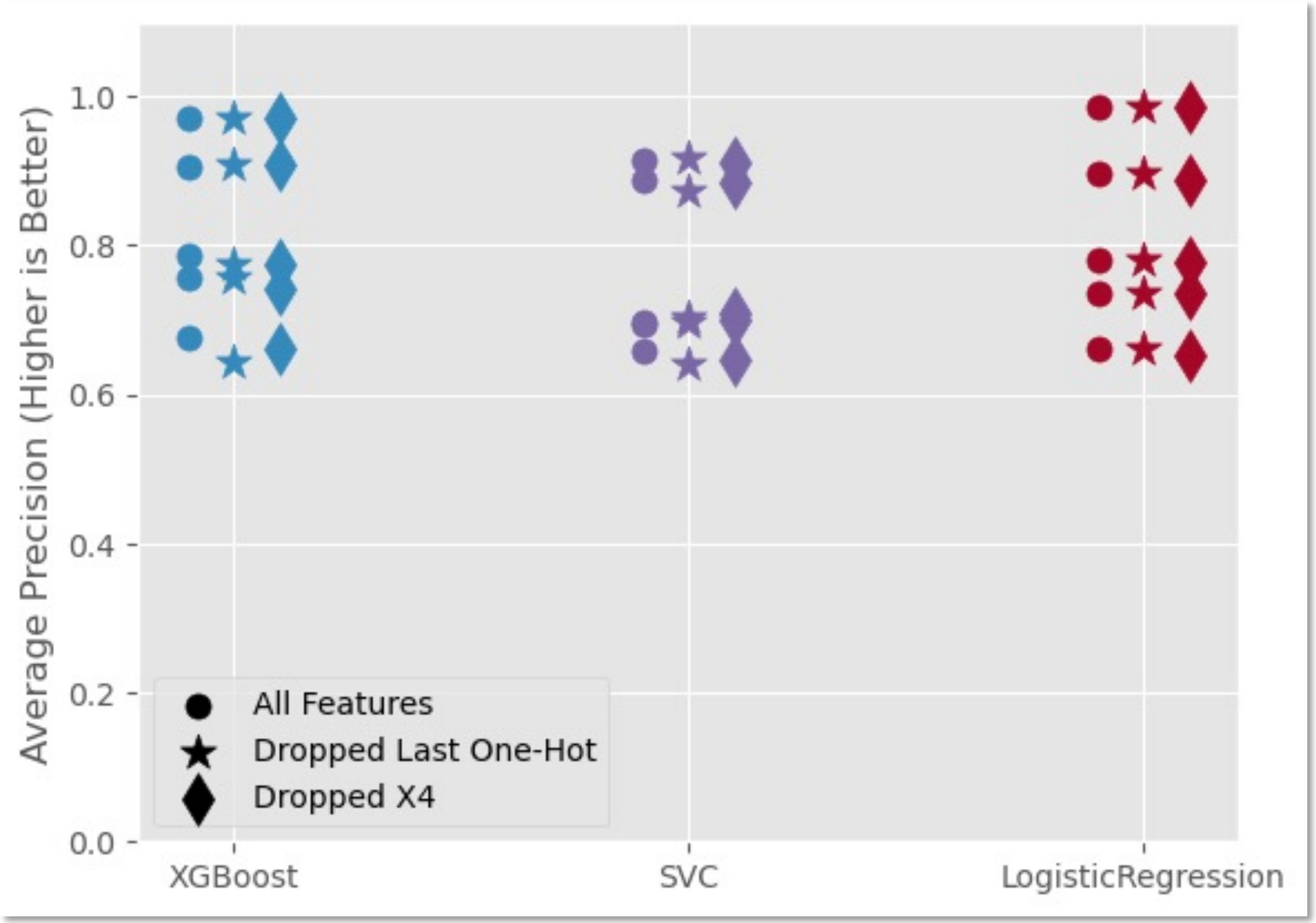
Yc
Boolean

Distinct	2
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	17.5 KiB

Classification



Classification



Conclusion 1/2

- Classification
  - LogReg or XGBoost do best (except at low thresholds)
  - XGBoost is the only one that stays above the baseline (“coin-toss”)
  - SVC isn’t doing very well
- Regression
  - Is really hard (target distribution is step-functiony), cf. horizontal stripes
- Overall, I’d go with XGBoost
  - It’s also “explainable”, so yay

Conclusion 2/2

- X4 seems redundant
  - Models don’t really change with/without

- Dropping one column from the one-hot encoded doesn’t matter
  - Models don’t change much with/without

What Now?

- Multi-Class Prediction?

- More Preprocessing
  - PCA? TDA?
  - Deal with noisy features? (X4, X5, X6)
    - → Fit Gaussian and take delta to be left with something interesting
    - → Or something more sophisticated (whitening / decorrelation)
    - Try dropping X6?

  - Drop one of one-hot? (Correlated)

- Pipeline in Sklearn

- Explainability