Dr. Csaba Burger, CFA (MNB)
Mihály Berndt (Clarity Consulting)

3rd IFC and Bank of Italy Workshop on “Data Science in Central Banking: Enhancing the access to and sharing of data”

ERROR SPOTTING WITH GRADIENT BOOSTING

Rome, 18th October 2023

MOTIVATION, BACKGROUND, SUMMARY OF RESULTS

Background

- MNB's commitment to high data quality
- Machine learning is suitable for large data volumes
- The role of ML in data quality checks is not yet standardized

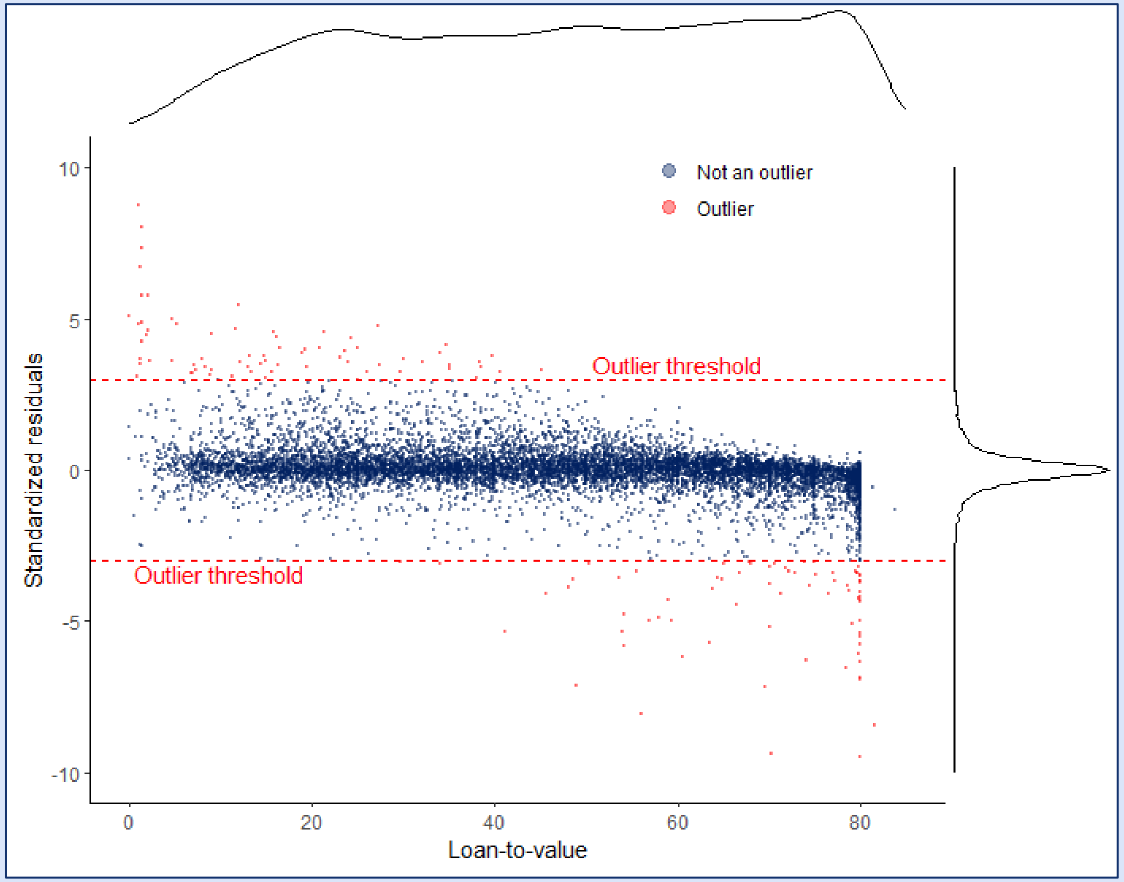
Results

- Un-labelled supervised learning can uncover relationships within the data
- State-of-the-art modelling techniques (XGBoost, Bayesian optimization)
- We present a few recommendations to flag potential data errors

Unlabelled supervised methods we use

1. Aggregated time series
2. Cross-sectional - granular
3. Granular time series

Residual plot in a model explaining a selected target variable

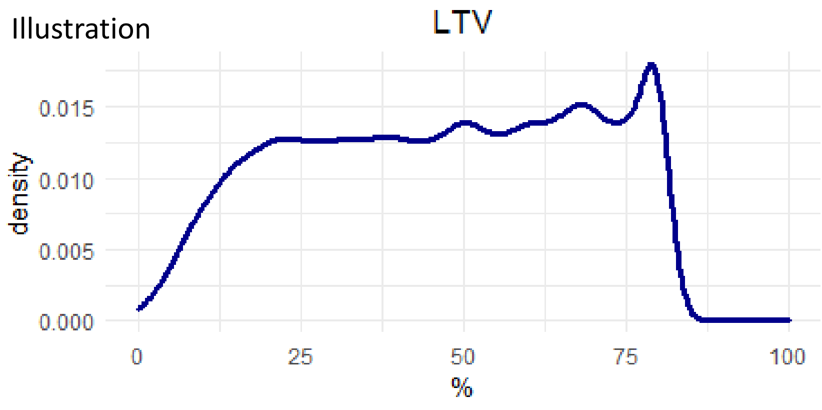


DATA

MNB LTV report

- First ranking mortgages with a start date after 1st Oct 2021
- Approx. 73 thousand lines
- 274 columns → 69 columns (high correlations, missing value share >= 20 percent)

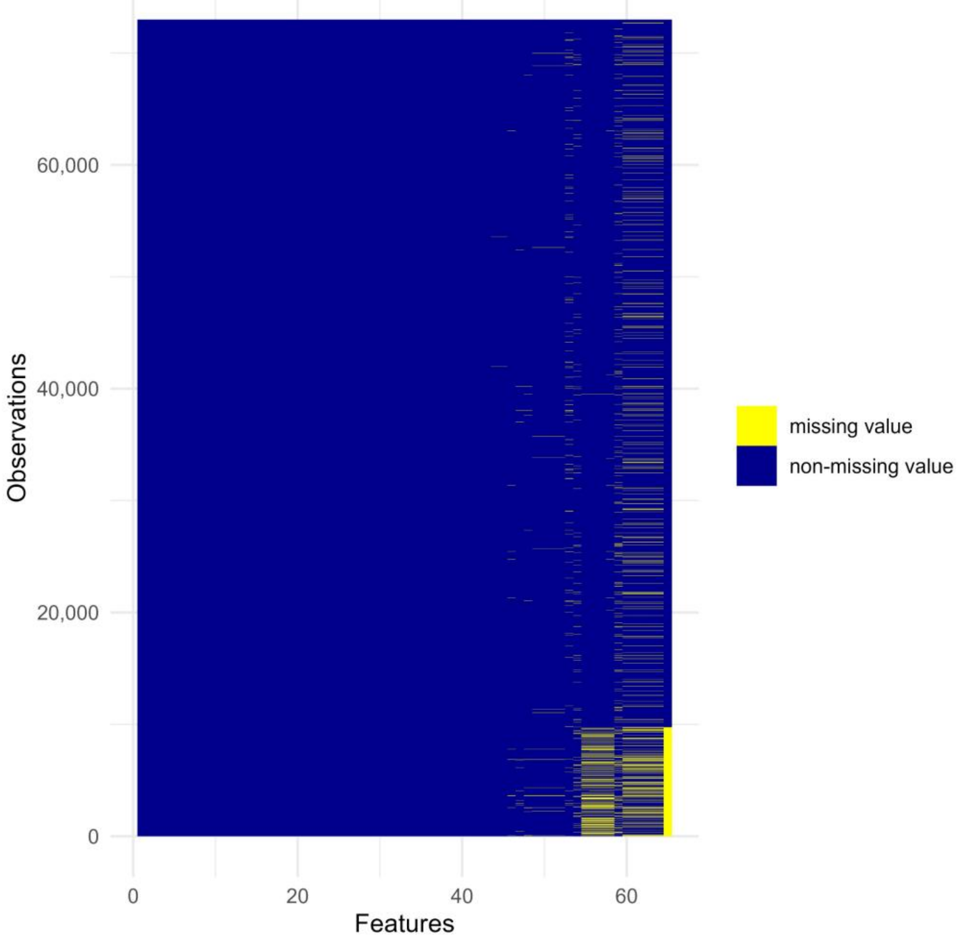
Illustration



Just a theory

$$\text{LTV} = \frac{\text{Loan amount}}{\text{Allocated collateral value}}$$

Missing values



XGBOOST GROWS DECISION TREES IN A SEQUENTIAL MANNER

Loss reduction calculation

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma.$$

Similarity scores based on:

- residual **direction**
- residual **magnitude**

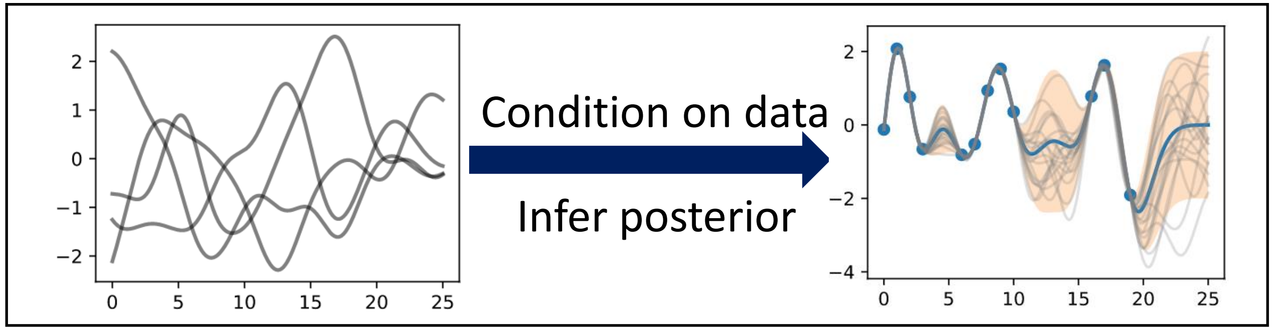
Many hyperparameters to optimize

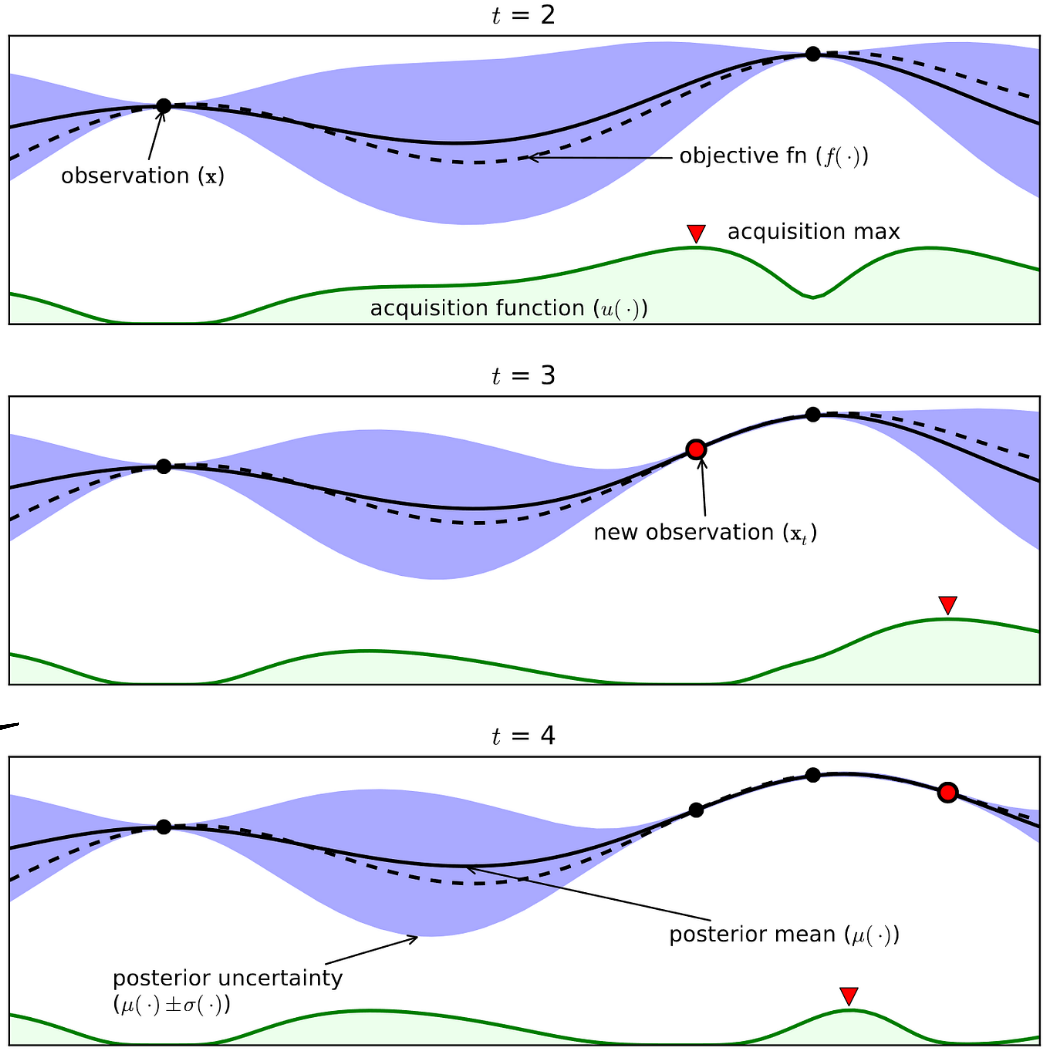
Sparsity-aware split finding

1. Visit only non-missing entries
2. Determine the best split and **default direction** for missing value based on the Similarity score above

GP: model of the objective function behaviour

- Train and test points are jointly distributed as multivariate normal
- Kernel encodes similarities between data points (shape of the prior)





How to determine new samples?

- Acquisition Function
- Exploration-explotation trade-off

MODELLING PIPELINE AND DESIGN QUESTIONS

Treatment of missing values → Treatment of rare values → Determining the loss function → Bayesian optimization → Interpreting the results

Question 1

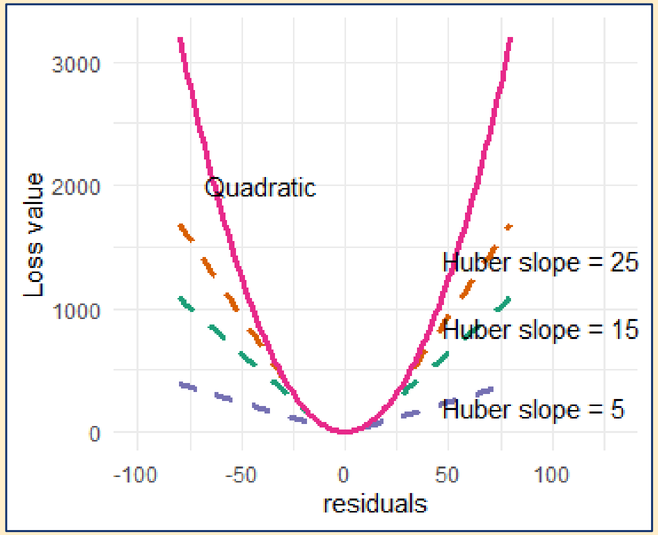
Ways to deal with missing values:

- Estimator (missRanger)
- Constant (unusual dummy)
- Xgboost's sparsity-aware split finding

Question 2

Loss function choice:

- Mitigate the impact of existing errors on finding the ground truth



Question 3

We assume:

- **If**: explanatory column 'B', is not independent from 'A'
- **and** data error distorts an explanatory variable 'A'
- Then B takes over from A

SYNTHETIC ERRORS
INSPIRATION FROM EXISTING ERRORS

Location	Description
Response variable	Values divided by 100.
Response variable	Values set to 80.
Response variable	Values were multiplied by a random value, drawn for each observation from U(0.4, 0.6) and U(1.2, 1.4)
Predictor (2nd most important)	Values set to 10 mln HUF
Predictor (2nd most important)	Values were multiplied by a random value, drawn for each observation from U(0.4, 0.6) and U(1.2, 1.4)

Errors in 5% of all observations, both in train and test sets

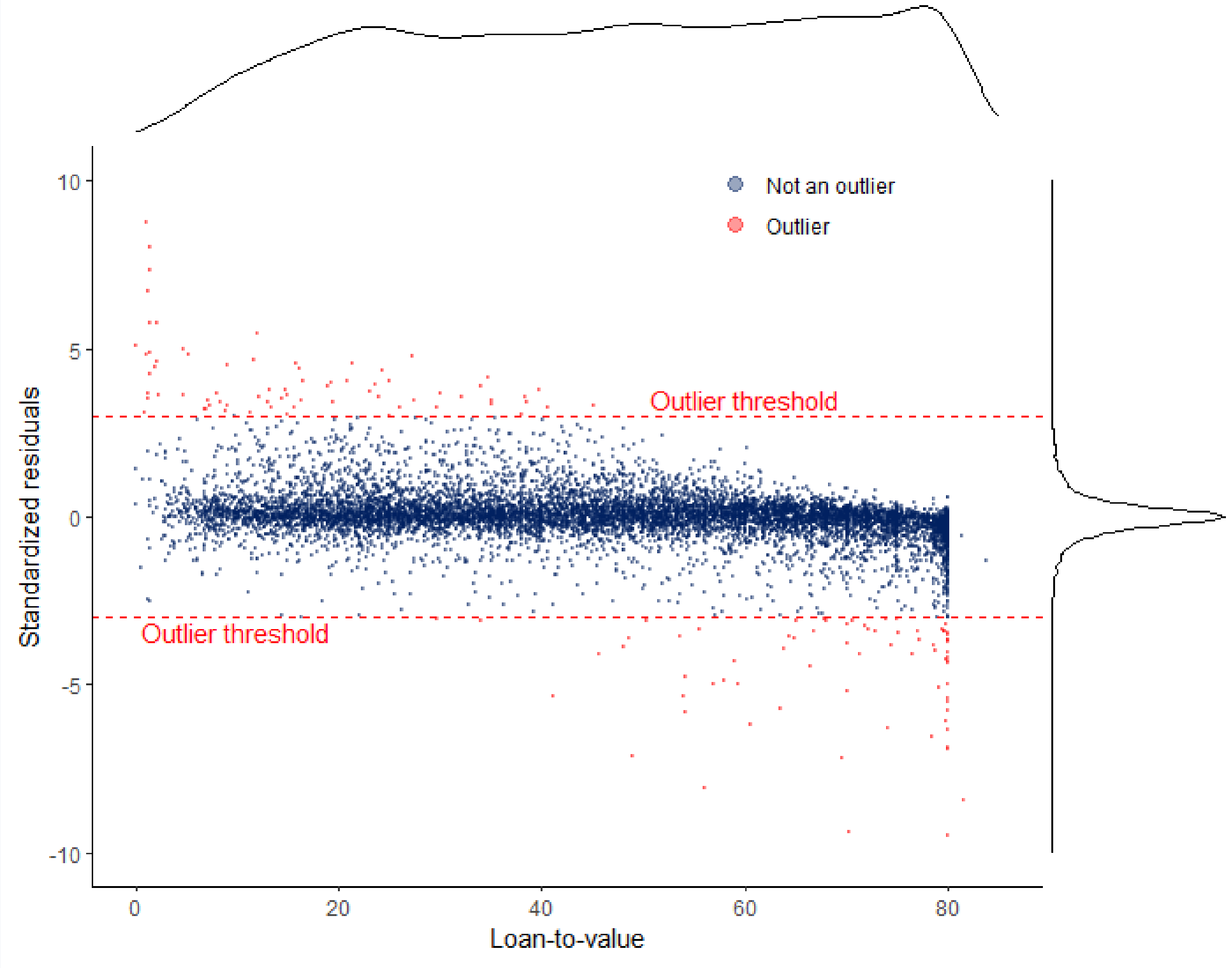
BASELINE MODEL PERFORMANCE
WITHOUT SYNTHETIC ERRORS

The baseline model

- missing values using a constant
- squared loss function
- **no synthetic errors**

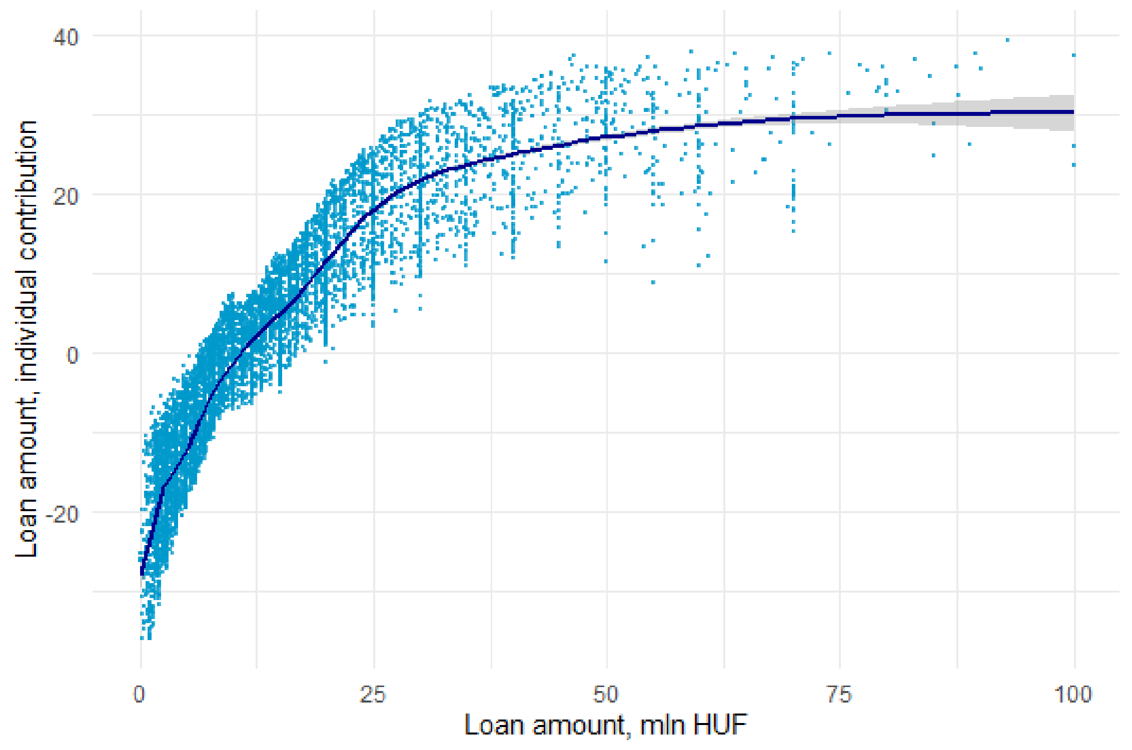
Model performance

- RMSE = 5.6 percent, MAE = 3.2 percent
- the share of outliers is 1.4 percent only (cutoff of *standardized* residuals of 3)
- The algorithm found intuitive errors (LTV as a fraction between 0 and 1)

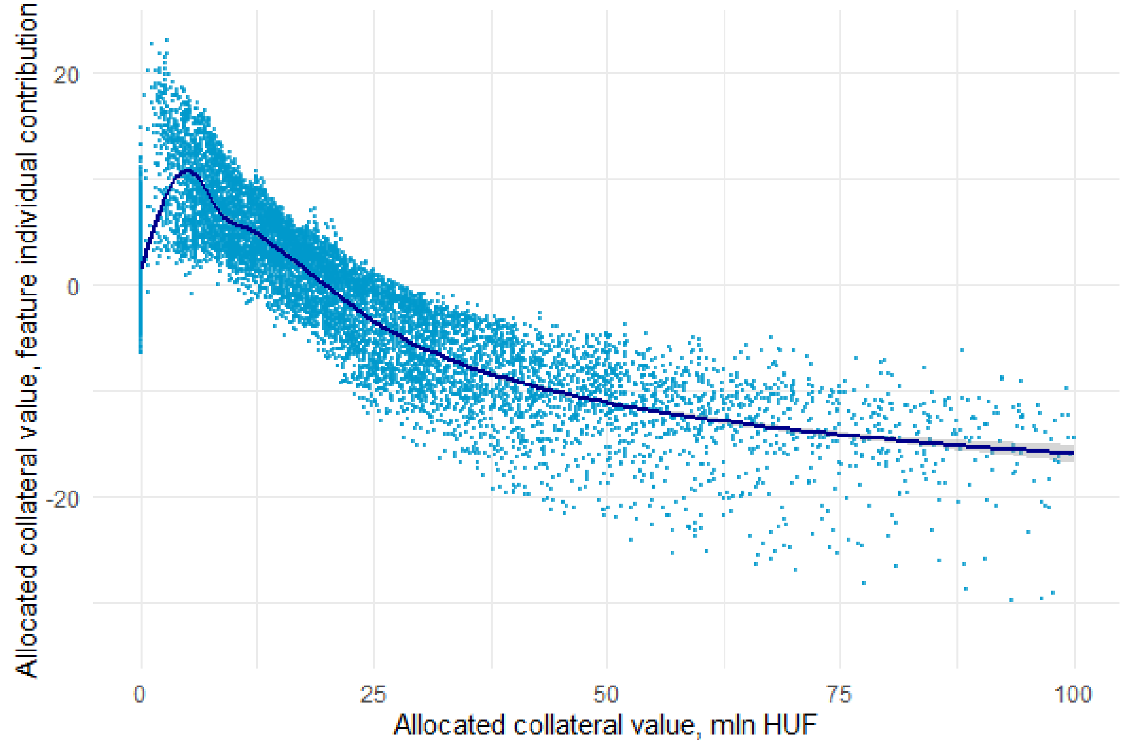


THE BASELINE MODEL – INDIVIDUAL FEATURE CONTRIBUTIONS (IFC)

IFC for Loan amount + a LOESS function



IFC for allocated collateral value + a LOESS function



$$\textbf{LTV} = \frac{\textbf{Loan amount}}{\textbf{Allocated collateral value}}$$

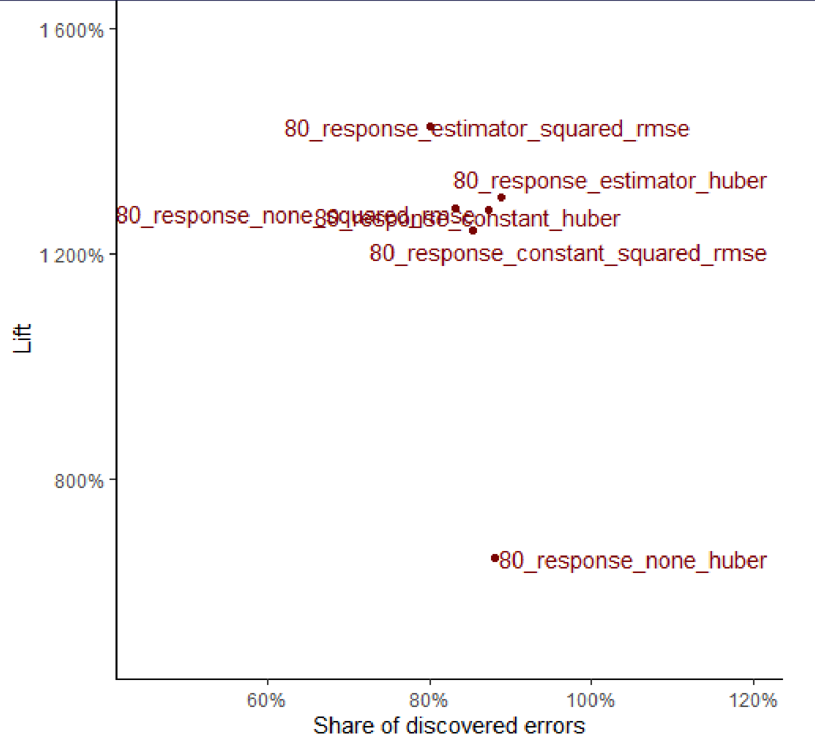
METRICS TO ASSESS MODEL PERFORMANCE

	Share of discovered errors	Lift value
Formula	$Disc.\,error\ sh. = \frac{Errors\ among\ outliers}{All\ errors}$	$Lift = \frac{Error\ share\ among\ outliers}{Error\ share\ in\ all\ data}$
Rationale	Did we find every synthetic error?	Am I any better off by looking at outliers than going through the raw data?

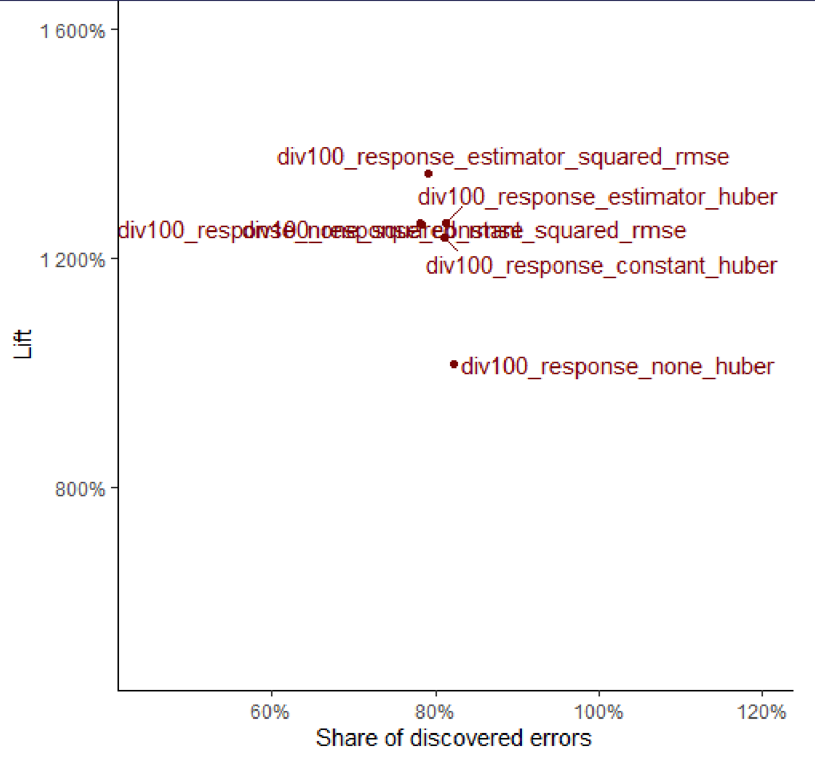
One metric is insufficient

HYPOTHESIS 1 (MISSING VALUE REPLACEMENT) AND HYPOTHESIS 2 (LOSS FUNCTION)

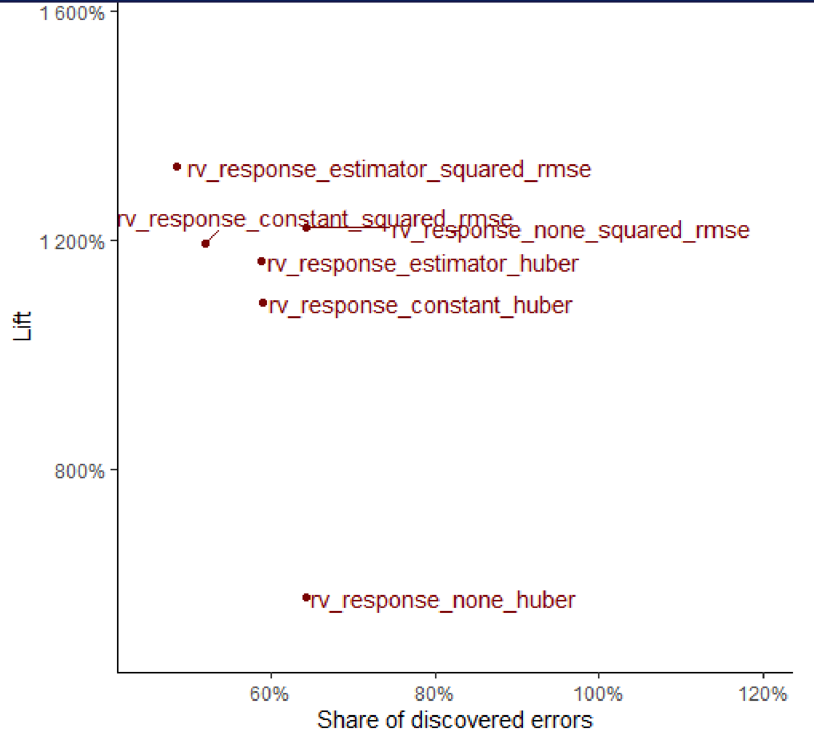
Error type
80



Error type
div 100



Error type
random value multiplication



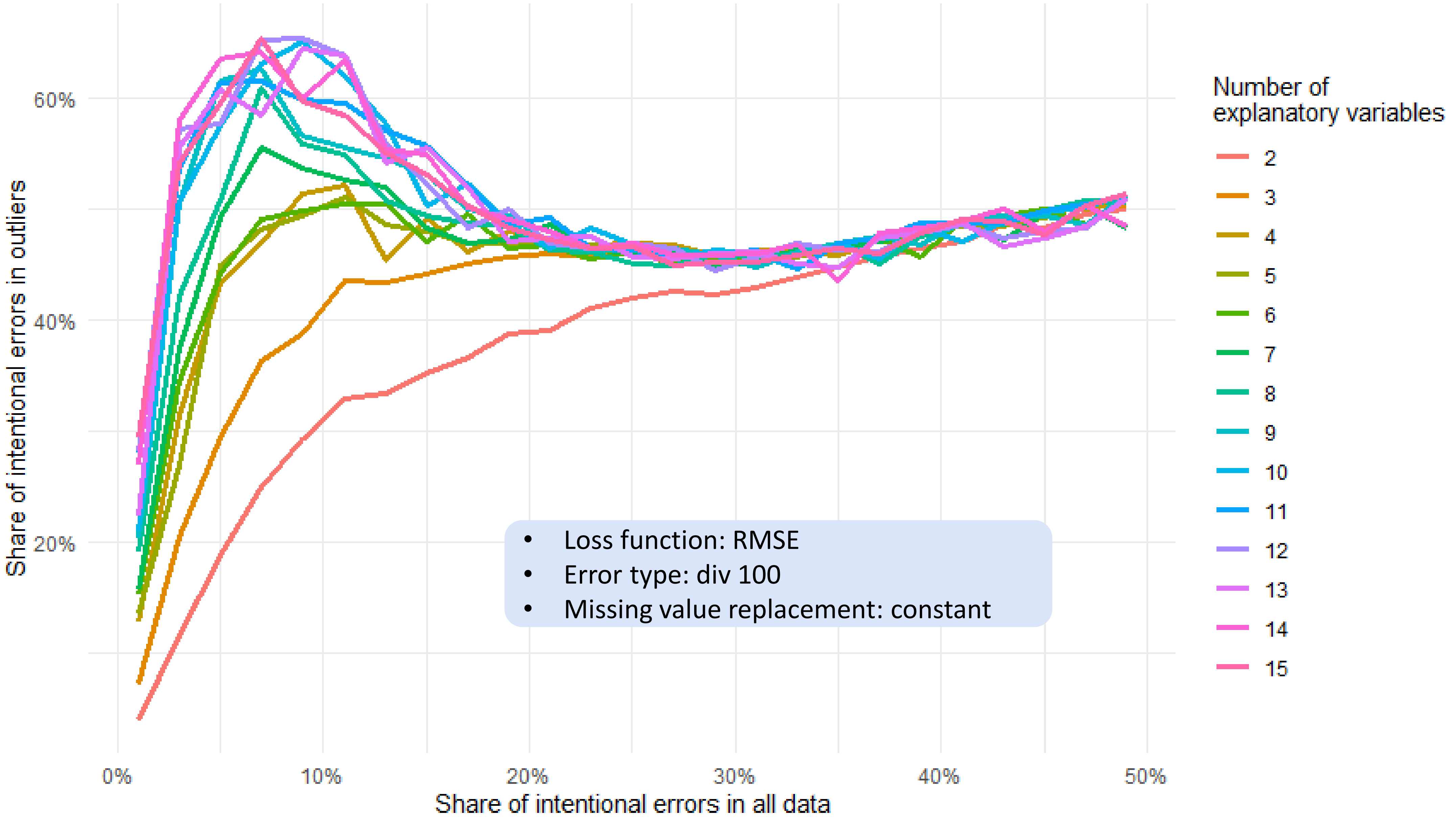
HYPOTHESIS 3 – ERROR IN ALLOCATED COLLATERAL VALUE

Error type	Missing value replacement	Loss function	Outliers as % of total	Share of discovered errors	Lift
10	none	Huber	3,5%	3,0%	0,87
10	estimator	Huber	3,0%	2,1%	0,70
10	none	rmse	1,5%	1,6%	1,05
10	constant	Huber	2,3%	1,3%	0,57
10	constant	rmse	1,3%	1,1%	0,83
10	estimator	rmse	1,1%	0,9%	0,83
rv	none	rmse	2,0%	10,7%	5,41
rv	estimator	Huber	4,5%	7,9%	1,75
rv	none	Huber	2,5%	5,8%	2,31
rv	constant	Huber	3,0%	4,3%	1,43
rv	constant	rmse	1,5%	2,1%	1,41
rv	estimator	rmse	1,3%	1,9%	1,51

Vs. 70-80 % when error in target

Vs. 10-12 when error in target

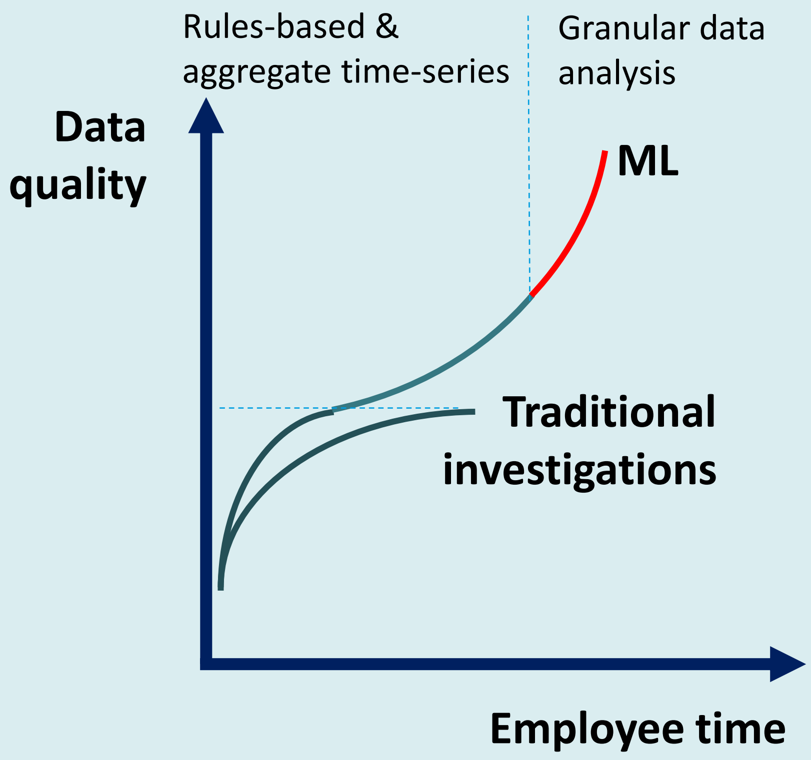
ABOVE NINE EXPLANATORY FEATURES AND AN ERROR SHARE OF AROUND 10 PERCENT EFFICIENCY STARTS TO DROP



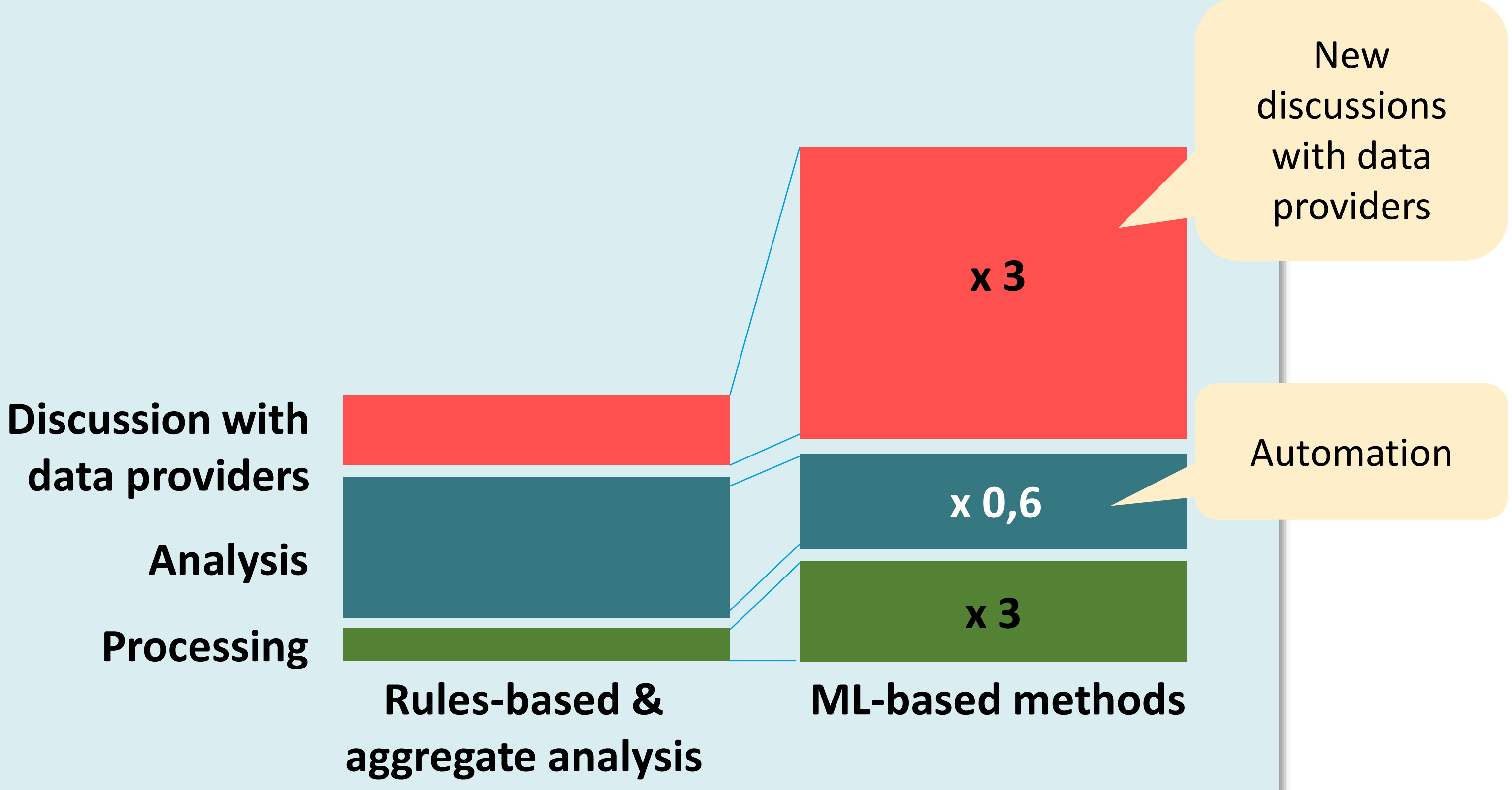
- Loss function: RMSE
- Error type: div 100
- Missing value replacement: constant

ML-BASED DATA QUALITY TESTS CREATE MORE WORK FOR OTHERS

Data quality explodes
but only if you work with it



Required human labour



SUMMARY

Findings recap

- A supervised learning algorithm to flag potential data errors
- The method successfully identifies synthetic errors
- It provides hints to their location
- We also analysed various steps during the preprocessing phase (missing values and loss function) which may improve performance

Implications

- Our results helps the data providers
- The ’last mile problem’ is still there: error flags do not provide interpretation
- Our results help modellers: model predictions may be used instead of actual values

Thank you for your attention!