

How Data Quality Affects Machine Learning Performance

A study using income and healthcare datasets

Author: Faizan Mohammed Yahya

Table of Contents

<i>Abstract</i>	3
<i>Introduction</i>	3
<i>Datasets</i>	4
<i>Data Quality Assessment</i>	4
<i>Methodology</i>	5
<i>Phase 1 Result</i>	7
<i>Phase 2 Results and Cross-Domain Comparison</i>	8
<i>Discussion and Limitations</i>	9
<i>Conclusion</i>	9
<i>References</i>	10

Abstract

Data Quality plays the most significant role in Machine Learning. As the machine learning model learns patterns from the input data, this data needs to be clean. The project, in this regard explores and investigates how a Machine Learning model can be affected by common data quality problems.

To do so, this project has been divided into two phases. The first phase uses the UCI Adult income dataset, and the second phase uses UCI chronic Kidney dataset. In both the phases, we used a consistent experimental design utilizing logistic regression.

In Phase 1, the class weighting increased the model recall from 0.603 to 0.844; however, it also decreased precision from 0.733 to 0.554.

Secondly, in phase 2, the outlier treatment corrected one error in the main test split, but this improvement did not appear in other tested splits. Therefore, this study finds that data cleaning does not always improve every aspect of data-quality. Its effect depends on the type of problem and the dataset.

Introduction

Machine-learning models learn patterns directly from the training, so problems such as missing values, duplicate records, outliers and class imbalance can distort what they learn. Cleaning is never risk-free, because an extreme value may be a genuine observation, and imputing a missing value can remove a useful signal. The effect of cleaning decision should therefore be tested rather than assumed.

This project examines how common data-quality interventions change the performance of a fixed classifier, using logistic regression and a consistent experimental design across two datasets from different domains, namely the UCI Adult Income dataset (for Phase 1) and the UCI Chronic Kidney Disease dataset (Phase 2). Interventions are tested individually and then combined in a full pipeline. The aim is not to build the highest performing model but to isolate how data-quality decisions affect a fixed model.

The main research question is: *How do common data-quality interventions change the performance of a fixed classification model, and are the effects similar in two different domains? The aim is to study how data quality interventions affect performance, not to build the highest-performing model.*

Datasets

Phase 1 is based on the Adult data of the UCI Machine Learning Repository (Becker & Kohavi, 1996). The analyzed file has 32,561 rows, 14 input features, and a binary target showing whether income is above \$50,000 per year. It was chosen since it consists of both numerical and categorical variables; it has natural missing values, a few repeating rows, and an unbalanced target.

Phase 2 is based on the UCI Chronic Kidney Disease data (Rubini et al., 2015). It has 400 records, 24 input features, and a CKD/not-CKD target. It represents a different domain and has a much more missing data. Both the datasets are freely available under the Creative Commons Attribution 4.0 license. The original data files were kept unchanged.

Data Quality Assessment

The initial action was the measurement of each problem. Field and cell counts of missingness were obtained. Duplicate counts are additional rows that have been precisely the same features and target, but they do not imply that rows represent the same person. The 1.5 x IQR rule had been used to identify numerical outliers. The values recorded by UCI were compared with categories. Measure of class balance were taken based on the target counts.

Table 1: Data Quality Assessment Metrics

Measure	Adult income	Kidney disease
Rows	32,561	400
Missing feature cells	4,262 (0.93%)	1,012 (10.54%)
Rows with missing features	2,399	242
Exact duplicate extras	24	0
Non-standard feature values	0	1
Numerical IQR flags	11,325	211
Positive class share	24.1%	62.5%

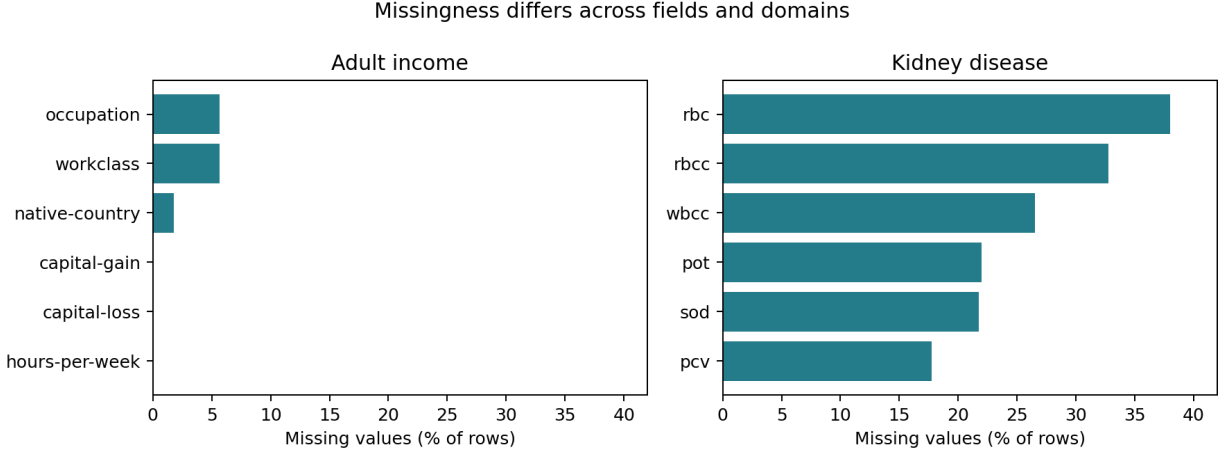


Figure 1: Fields with highest missing values in both datasets

Mostly, missing entries in adults are in work class, occupation, and native-country, The minority class is the positive income class. In the kidney dataset, there are a lot of missing values in numerous fields, such as red blood cell measurements, sodium, and potassium. The kidney dataset does not contain any exact duplicate rows. One of the values in diabetes has an additional white space and two target labels. Also, it has an additional trailing white space. All the experiments have standardized target labels, so the result would be the same.

The fact that there are large numbers of IQR flags does not imply that the datasets have thousands of errors that certain. Examples are that a long working week or high medical measurements are not usual but real. Outlier treatment has instead been tested rather than being presumed as being correct.

Methodology

The data has been split into training and test sets with a ratio of 80/20 with `random_state = 42`. The records of identical features were stored on the same of the split to minimize leakage of duplicates. The learning of all imputation values, category encoding, scaling values, outlier limits, and class weights was done only on the training data. This excludes the test data from providing information to the model (scikit-learn developers, n.d. -a).

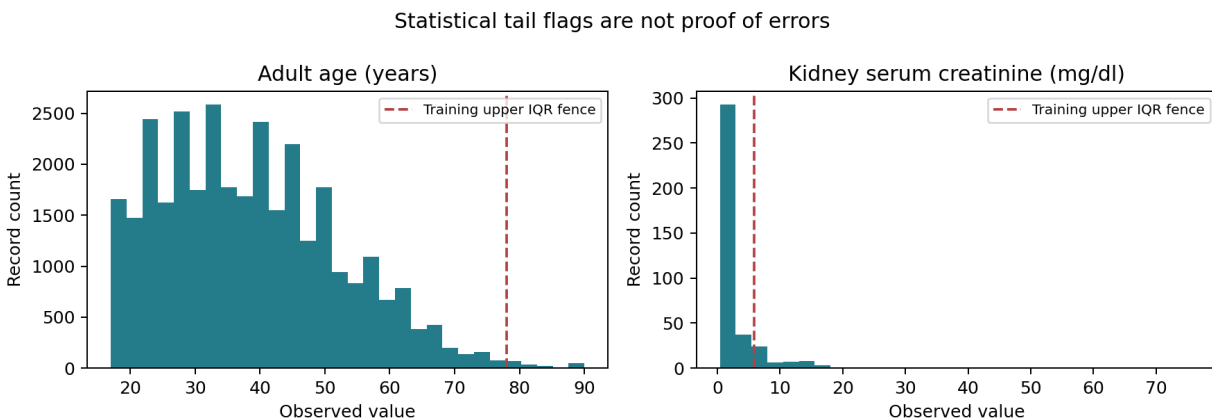
Logistic regression was chosen due to being a standard and interpretable classifier. Numerical features were transformed into standardized and categorical features into one-hot encoded features. The model was based on the liblinear solver, $C=1$, a classification threshold of 0.5, and a fixed random state. The measures were mainly precision, recall, and F1, and the secondary measures were accuracy and ROC-AUC

Table 1: Experiments used in both project phases

Experiment	Change tested
E0	Baseline with simple placeholders required to run the model
E1	Training median for missing numbers and mode for missing categories
E2	Remove exact duplicate rows from the training set.
E3	Clip numerical tails using training-set IQR limits
E4	Standardise documented categories and invalid values.
E5	Use balanced class weights during model training.
E6	Combine E1, E2, E4 and E5; retain plausible extreme values

Simply encoding and the use of missing-value placeholders are also required on the baseline since the logistic regression does not know how to directly work with text and empty numerical values. E1 compares a more informed imputation rule with those placeholders. E5 does not make any artificial records; it alters the significance attributed to errors of each training class. The test scores were compared with the definition of E6.

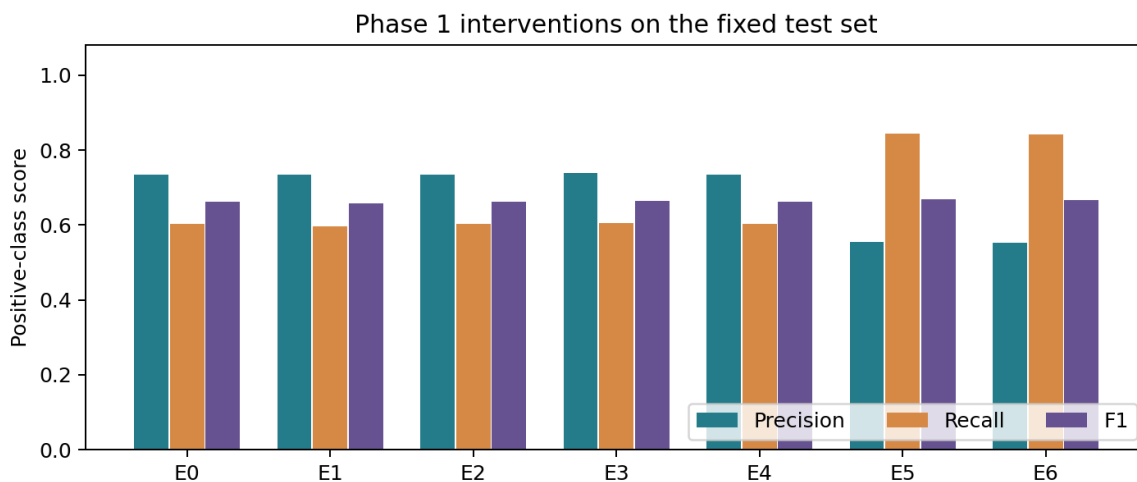
Random states 43 and 44 were also re-experimented in the entire experiment. Though these additional splits are merely sensitivity tests. They also assist in demonstrating that a little increase in contingent upon single train/test distribution, however; they are not a statistical significance test.

**Figure 2: Statistical outlier limits for one field from each domain**

Phase 1 Result

Table 2: results of Adult income data after main test split

Test	Accuracy	Precision	Recall	F1	ROC-AUC
E0	0.852	0.733	0.603	0.662	0.903
E1	0.851	0.734	0.596	0.658	0.901
E2	0.852	0.734	0.603	0.662	0.903
E3	0.853	0.739	0.605	0.665	0.905
E4	0.852	0.733	0.603	0.662	0.903
E5	0.799	0.554	0.844	0.669	0.903
E6	0.798	0.552	0.842	0.667	0.901



The baseline got $F1 = 0.662$. F1 was slightly decreased to 0.658 due to median/mode imputation. Deleting 22 duplicate rows of training resulted in a very minimal change. The standardization of categories did not matter since no values of Adult features violated the rules chosen to be documented.

The most apparent change was caused by class weighting (Experiment 5). The recall rose to 0.844, compared to 0.603, and thus the model identified more cases of high income. Precision fell from 0.733 to 0.554, and accuracy fell from 0.852 to 0.799. It is not a mere enhancement but a trade-off. The combined E6 pipeline maintained a similar trend and failed to do better than E5 on F1.

Phase 2 Results and Cross-Domain Comparison

Table 3: Kidney disease results on the main fixed test split

Test	Accuracy	Precision	Recall	F1	ROC-AUC
E0	0.988	1.000	0.980	0.990	0.980
E1	0.975	1.000	0.960	0.980	0.999
E2	0.988	1.000	0.980	0.990	0.980
E3	1.000	1.000	1.000	1.000	1.000
E4	0.988	1.000	0.980	0.990	0.980
E5	0.988	1.000	0.980	0.990	0.980
E6	0.963	1.000	0.940	0.969	0.999

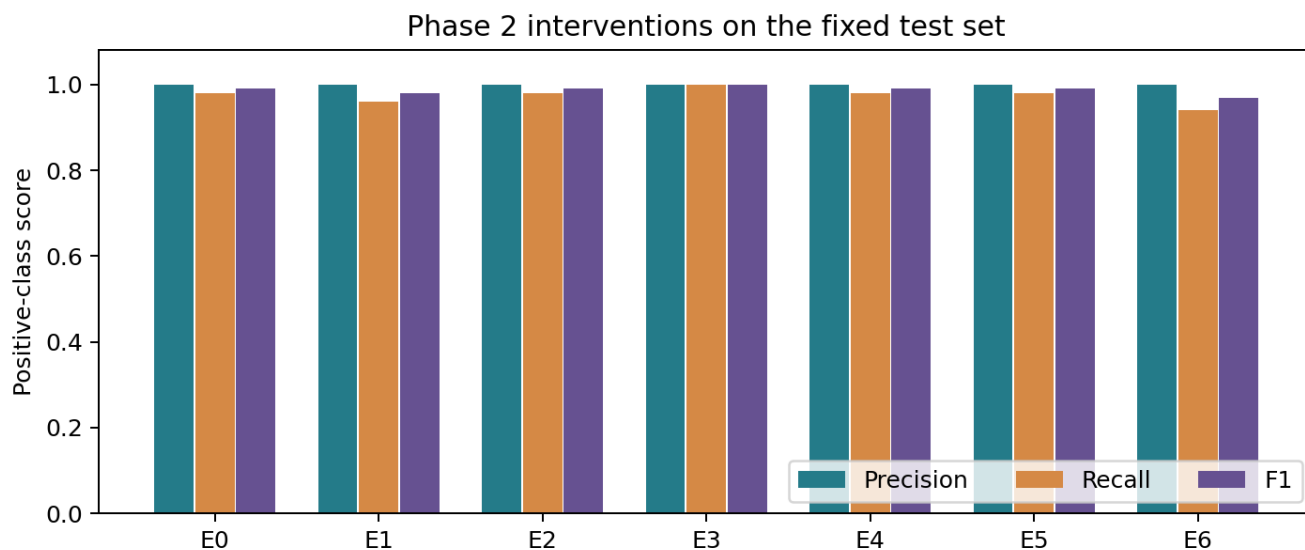


Figure 3: Precision, recall, and F1 for the kidney experiments

The kidney baseline was already quite high, and $F1 = 0.990$ on 80 test records. E1 increased ROC-AUC from 0.980 to 0.999, but F1 decreased because the model missed one additional CKD case at the 0.5 classification threshold. E2 did not impact anything since this dataset did not contain any exact duplicates. A single baseline error in the main split was corrected by E3, and

F1 = 1.000 was obtained. But its F1 gain was zero with random states 43 and 44. Thus, the overall finding is not a strong indication that clipping would be necessarily beneficial.

E4 standardised one non-standard categorical value in the training data, but this did not change the model's predictions. Class weighting did not change the main test predictions because the model was already separating the CKD and non-CKD cases very well. The combined pipeline decreased F1 to 0.969 because the model missed more positive CKD cases.

The two domains thus acted differently. Kidney data had more missing with higher baseline performance. Class weighting had more effect on adult, whereas small changes in kidney varied according to the test split. Profiling checks can be reused across domains, although it requires domain knowledge to decide whether to clean or retain a value or not.

Discussion and Limitations

In the Adult study, class weighting caused the model to predict more positive cases, but it also produced more false positives. An appropriate decision would also be based on the error that is more important in the desired application. Treating missing values did not always assist. An additional missing category can occasionally contain data on the manner in which the data was gathered.

Additionally, there are some important limitations to discuss about this study. The kidney dataset is small; thus, a single prediction can significantly alter the score. The Adult data is historical and has sensitive demographic fields. There is only a single model that was examined, as opposed to one such study where 19 different models were used (Mohammed et al. 2025), and there are also restricted additional train/test splits. IQR limits can be used to identify unusual values, but they cannot be used to demonstrate their incorrectness. The research also fails to test the fairness among demographic groups or clinical performance in a new hospital.

The results show what changed after each preprocessing intervention, but they do not prove that the same intervention would cause the same effect in other datasets. The kidney model's high accuracy does not mean that it is ready to diagnose real patients. The high score may partly reflect how clearly the measurements in this small dataset separate the provided class labels.

Conclusion

The two machine-learning tasks were impacted by data-quality interventions differently. The best Adult performance was a recall-precision trade-off following class weighting. The outlier clipping in the kidney experiment increased a single main-split prediction, but was not replicated in other splits. Imputation and combined cleaning did not prove to be very useful in all cases.

The general finding is that data quality is to be quantified cautiously and any cleaning decision is to be experimented with instead of presumed to enhance a model.

References

Becker, B., & Kohavi, R. (1996). Adult [Data set]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5XW20>

*Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12, 2825-2830.*

Rubini, L., Soundarapandian, P., & Eswaran, P. (2015). Chronic Kidney Disease [Data set]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5G020>

scikit-learn developers. (n.d.-a). Common pitfalls and recommended practices. https://scikit-learn.org/stable/common_pitfalls.html

scikit-learn developers. (n.d.-b). LogisticRegression documentation. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html