# Enhancing Data to Boost Machine Learning Model Performance

December 19, 2022

## Introduction

For the last two decades the dominant approach to artificial intelligence (AI) has been model-centric, with the aim of developing a machine learning model capable of dealing with data gaps by iteratively improving the model until it reaches optimum performance while maintaining the data as it is. However, to further improve performance, the model's training data must be optimized, and in 2021, a data-centric approach to optimizing AI models began to emerge[1], an approach which is based on keeping the model or code constant while iteratively improving the data's quality.

In the last year, we also have witnessed the rise of generative AI. Much of the enthusiasm surrounding it dealt with content such as images, videos, email, social media, voice recordings, and program code, which were utilized to generate new content, translations, answers to queries, sentiment analysis, summaries, and even videos. But exciting work is happening in the structured data scene as well, where generative models are being used for data augmentation, balancing, imputation, cleansing, and sharing. Given its demonstrated capabilities, it is only natural to consider leveraging generative AI's strengths to improve data quality, as generative AI holds the promise of transforming the data we collect into the data we want.

Before demonstrating how machine learning model performance can be boosted, it is worth discussing the relationship between a machine learning task, the aim of which is to make a prediction, and the data available for performing that task. There is a direct and essential connection between a machine learning task and the data used to perform that task. Data may be imbalanced, incomplete, mislabeled, noisy, biased, or flawed in some other way. As a result, predictions may be inaccurate at best and harmful at worst; models that produce racial, gender, or geographic bias potentially expose organizations to devastating consequences. If instead we can optimize the data, preventing such flaws before model training ever begins, machine learning predictions will be more effective, efficient, and accurate.

In the classification task the model assigns data to categories, while in the regression task the model predicts the value of the label from a set of related features. In the regression task the label can be any real value and is not limited to a closed set of values, unlike the classification task.

In binary classification tasks such as defect prediction, fraud detection, and disease detection, only two possible prediction outcomes exist. These tasks are often imbalanced, meaning that most predictions fall into the majority class, and only rare cases fall into the minority class. Sometimes the minority class is the more interesting case, particularly given the importance of identifying instances in the minority class in order to ensure that most actual fraud, defects, or ill patients are identified, even at the cost of false alarms. However, it is usually more difficult to predict the minority class, since it is not sufficiently represented in the source data. Balancing

techniques are commonly used to deal with imbalanced data, either by undersampling the majority class, oversampling the minority class, or both.

In multiclass target classification tasks, there are more than two possible prediction outcomes. For example, predicting the type or level of an examined phenomenon/customer/disease. Here too, we often face imbalanced classes, so majority classes might be easier to predict than minority classes.

In contrast to the two categorical tasks mentioned above, in numeric prediction tasks the prediction outcome is a continuous number, such as a score, age, price, or amount. Not all values in the permitted range may be represented and some values may be represented more than others. Augmentation and balancing can be used to optimize data for these tasks, yielding richer data with more potential prediction values.

In this document, we demonstrate the power of our data-centric generative AI-based approach and its ability to produce more accurate prediction models simply by optimizing the training data.

## Our Data Enhancement Process

In our data enhancement process, a variational autoencoder (VAE) is used to transform the input data into a lower-dimensional feature space, where new representations are generated using techniques like crossover and SMOTE. These are later decoded back to the original feature space and serve as augmented records for the source data. Grid search is performed on a validation set to optimize various process-related parameters like the augmentation and balancing ratios, given a specific evaluation metric.

## Experiment

We evaluated our data enhancement process on 60 small- to medium-sized datasets (datasets with hundreds to tens of thousands of records), of which 20 were related to a binary target classification task, 20 to a multiclass target classification task, and 20 to a numeric target regression task. The datasets and their metadata are presented in Table 1.

*Table 1: Input datasets' metadata*

| Type of Target | ID | Data | Fields | Categorical Fields | Numeric Fields | Majority Ratio | Target Classes | Records | Augmented Records | New Majority Ratio |
|---|---|---|---|---|---|---|---|---|---|---|
| Binary | 1 | stroke | 11 | 8 | 3 | 95% | 2 | 5,110 | 446% | 58% |
| | 2 | system | 16 | 5 | 11 | 95% | 2 | 19,084 | 594% | 56% |
| | 3 | champions | 11 | 5 | 6 | 96% | 2 | 720 | 110% | 83% |
| | 4 | beauty | 10 | 8 | 2 | 93% | 2 | 1,260 | 602% | 56% |
| | 5 | covid | 110 | 25 | 32 | 90% | 2 | 5,644 | 504% | 56% |
| | 6 | sylvine | 21 | 1 | 20 | 50% | 2 | 5,124 | 117% | 50% |
| | 7 | sick | 30 | 14 | 6 | 94% | 2 | 3,772 | 185% | 69% |
| | 8 | qsar_biodeg | 42 | 12 | 29 | 66% | 2 | 1,055 | 655% | 52% |
| | 9 | page_blocks | 11 | 1 | 10 | 90% | 2 | 5,472 | 109% | 79% |
| | 10 | mozilla4 | 6 | 2 | 4 | 67% | 2 | 15,545 | 85% | 66% |
| | 11 | magic_telescope | 12 | 1 | 10 | 65% | 2 | 19,020 | 81% | 65% |
| | 12 | kr_vs_kp | 37 | 32 | 0 | 52% | 2 | 3,196 | 118% | 52% |
| | 13 | hmeq_p | 15 | 5 | 9 | 80% | 2 | 5,960 | 607% | 54% |
| | 14 | compas_two | 14 | 12 | 2 | 53% | 2 | 5,278 | 609% | 50% |
| | 15 | ada | 49 | 35 | 6 | 75% | 2 | 4,147 | 179% | 61% |
| | 16 | mimic_icu | 49 | 11 | 38 | 86% | 2 | 1,177 | 182% | 66% |
| | 17 | titanic | 12 | 6 | 3 | 62% | 2 | 891 | 275% | 53% |
| | 18 | loan_data | 28 | 4 | 10 | 84% | 2 | 19,156 | 365% | 54% |
| | 19 | hospital | 18 | 6 | 12 | 83% | 2 | 5,956 | 426% | 56% |
| | 20 | page_blocks | 11 | 1 | 10 | 90% | 2 | 5,472 | 268% | 62% |
| Multiclass | 1 | glass | 10 | 1 | 9 | 36% | 6 | 214 | 171% | 17% |
| | 2 | letter | 17 | 1 | 16 | 4% | 26 | 20,000 | 85% | 4% |
| | 3 | mfeat_zernike | 48 | 1 | 47 | 10% | 10 | 2,000 | 80% | 10% |
| | 4 | satimage | 37 | 1 | 36 | 24% | 6 | 6,430 | 114% | 17% |
| | 5 | soybean | 36 | 36 | 0 | 13% | 19 | 683 | 206% | 5% |
| | 6 | thyroid | 31 | 15 | 8 | 74% | 21 | 9,172 | 1240% | 5% |
| | 7 | led | 8 | 8 | 0 | 12% | 10 | 500 | 92% | 10% |
| | 8 | fetal_health | 22 | 7 | 14 | 78% | 3 | 2,126 | 187% | 33% |
| | 9 | autos | 26 | 11 | 14 | 33% | 5 | 205 | 129% | 20% |
| | 10 | ecoli | 8 | 2 | 5 | 44% | 5 | 336 | 173% | 22% |
| | 11 | segment | 11 | 8 | 2 | 67% | 4 | 8,068 | 213% | 25% |
| | 12 | sky_server | 18 | 2 | 12 | 74% | 3 | 10,000 | 178% | 33% |
| | 13 | cirrhosis | 20 | 8 | 6 | 56% | 4 | 418 | 180% | 25% |
| | 14 | fifa19 | 18 | 6 | 10 | 12% | 25 | 18,207 | 237% | 4% |
| | 15 | body_per | 12 | 2 | 10 | 60% | 4 | 13,393 | 191% | 25% |
| | 16 | multi_run | 17 | 3 | 14 | 64% | 5 | 21,726 | 258% | 20% |
| | 17 | microbes | 25 | 1 | 24 | 24% | 10 | 30,527 | 194% | 10% |
| | 18 | tablet | 20 | 10 | 10 | 34% | 9 | 2,000 | 243% | 11% |
| | 19 | crystal_structure | 18 | 6 | 10 | 61% | 5 | 5,329 | 244% | 20% |
| | 20 | accidents | 12 | 3 | 8 | 30% | 4 | 10,000 | 98% | 25% |
| Numeric | 1 | auto_mpg | 9 | 2 | 6 | | | 398 | 187% | |
| | 2 | bike_sharing | 16 | 6 | 10 | | | 730 | 173% | |
| | 3 | car_price | 26 | 9 | 14 | | | 205 | 118% | |
| | 4 | forest_fires | 13 | 4 | 8 | | | 517 | 116% | |
| | 5 | gemstone | 10 | 3 | 7 | | | 26,967 | 206% | |
| | 6 | jobs | 11 | 5 | 3 | | | 500 | 242% | |
| | 7 | motorcycle | 7 | 2 | 4 | | | 1,061 | 140% | |
| | 8 | profit | 5 | 1 | 4 | | | 50 | 120% | |
| | 9 | real_estate | 8 | 1 | 7 | | | 414 | 369% | |
| | 10 | sales_small | 5 | 1 | 4 | | | 4,572 | 302% | |
| | 11 | student_per | 33 | 29 | 4 | | | 649 | 163% | |
| | 12 | walmart | 8 | 1 | 7 | | | 6,435 | 272% | |
| | 13 | car_seats | 12 | 4 | 8 | | | 400 | 159% | |
| | 14 | insurance | 7 | 4 | 3 | | | 348 | 190% | |
| | 15 | bigmart | 12 | 8 | 4 | | | 8,523 | 292% | |
| | 16 | bodyfat | 15 | 0 | 15 | | | 252 | 152% | |
| | 17 | cellphone | 14 | 5 | 9 | | | 161 | 246% | |
| | 18 | house_rent | 12 | 8 | 3 | | | 4,746 | 126% | |
| | 19 | uso | 81 | 8 | 73 | | | 1,718 | 242% | |
| | 20 | pesticides | 13 | 3 | 10 | | | 8,760 | 268% | |

When performing the data enhancement process, the F1 score evaluation metric is used to optimize the free parameters when the data is related to a task with a categorical target, and the RMSE metric is used for the same purpose for data that relates to a task with a numeric target.

The metrics used to evaluate and compare the prediction models' performance when trained on the original data versus the optimized data are described in Table 2.

For datasets related to tasks with categorical targets (binary or multiclass classification) we measure the difference (improvement) in the following metrics: the F1 score, recall, precision, balanced accuracy, and ROC AUC. So given score A for a model trained on the original training data and score B for a model trained on the optimized training data, we calculate the prediction improvement score C, which is the increased score obtained by the optimized data: C = B – A.

For datasets related to tasks with numeric targets, we measure the decrease in the following three metrics: MSE, RMSE, and MAE. In this case, scores A and B represent the models' error, and we calculate the error reduction score C, which is the reduction in error from the original error: C = (A-B)/A.

*Table 2: The evaluation metrics used to evaluate our data enhancement process*

| **Evaluation Metric** | **Target Type** | **Description** |
|---|---|---|
| F1 Score | Categorical | Interpreted as the harmonic mean of the precision and recall, where an F1 score ranges between zero (poor performance) and one (best performance). The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is: F1 = 2 * (precision * recall) / (precision + recall). The F1 score is also known as the balanced F-score or F-measure. |
| Recall | Categorical | The ratio TP / (TP + FN), where TP is the number of true positives, and FN is the number of false negatives; the best value is one, and the worst value is zero. Intuitively, the recall represents the classifier's ability to identify all of the positive samples. |
| Precision | Categorical | The ratio TP / (TP + FP), where TP is the number of true positives, and FP is the number of false positives; the best value is one, and the worst value is zero. Intuitively, the precision represents the classifier's ability to avoid assigning a positive label to a negative sample. |
| Balanced Accuracy | Categorical | The average recall obtained by each class, where the best value is one, and the worst value is zero; this metric is suitable for imbalanced datasets. |
| ROC AUC | Categorical | The area under the curve of the receiver operating characteristic (ROC AUC) based on the prediction scores. |
| MSE | Numeric | Mean squared error regression loss. |
| RMSE | Numeric | Root mean square error, which is the standard deviation of the residuals (prediction errors). |
| MAE | Numeric | Mean absolute error regression loss. |

In our evaluation of the data enhancement process, five-fold cross-validation was performed on six prediction models (algorithms).
The evaluation flow for a single dataset is as follows:

- For each train-test split in the five-fold cross-validation:
  - A generative VAE model is trained based on the original training data, and the trained VAE model is then used to generate optimized training data.
  - The six algorithms listed in Table 3 are used to:
    - Train model A with the original training data
    - Train model B with the optimized training data
- For each of the relevant evaluation metrics:
  - The prediction improvement/error reduction score (for categorical/numeric targets respectively) is calculated based on the models' performance on the test data.
  - The average prediction improvement/error reduction score is calculated for each algorithm, and later the average prediction improvement/error reduction score for all the examined algorithms is calculated.

Finally, the average prediction improvement/error reduction score according to each of the relevant evaluation metrics is calculated for the aggregated scores obtained in the five train-test splits.

*Table 3: Algorithms used to evaluate the data enhancement process*

| **Algorithm** |
|---|
| CatBoost |
| GBM |
| LGBM |
| Logistic Regression |
| Random Forest |
| XGBoost |

## Results

Figure 1 provides a summary of the results.

Our results overwhelmingly demonstrate the benefit of optimizing the training data on model performance and the ability of the Datomize-enhanced datasets to produce accurate prediction models. Each of the datasets evaluated focused on a specific type of prediction task (binary target, multiclass target, or numeric target prediction), and in each case, the optimized training data contributed to improved model performance. Note that the results presented represent the average results obtained by all of the algorithms examined based on all five folds.

For 80% of the binary target tasks, the F1 score increased by 12.58% when the optimized training data was used; for 85% of these tasks, the recall increased by 20.79%; for 65% of them, the balanced accuracy score increased by 5%; for 45%, the precision increased by 2.42%; and for 25% of the tasks, the ROC AUC increased slightly by 0.75%.

For 55% of the multiclass tasks, the F1 score increased by 4% when the optimized training data was used; for 45% of these tasks, the recall increased by 9%; for 75% of them, the balanced accuracy increased by 5%; for 70%, the precision increased by 2%; and for 20% of the tasks, the ROC AUC increased slightly by 0.27%.

For 85% of the numeric target tasks, the RMSE decreased by 42% and the MSE decreased by 63% when the optimized training data was used; and for 80% of these tasks, the MAE decreased by 38%.

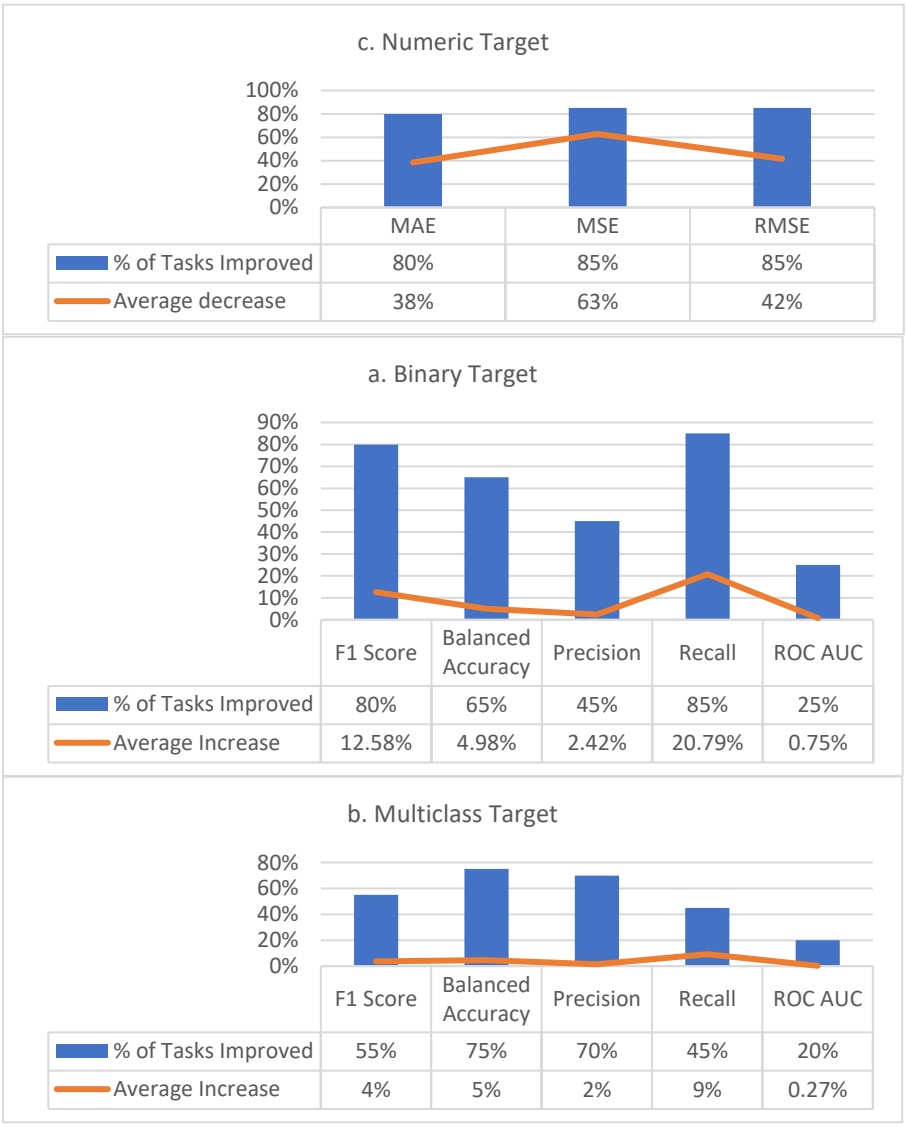
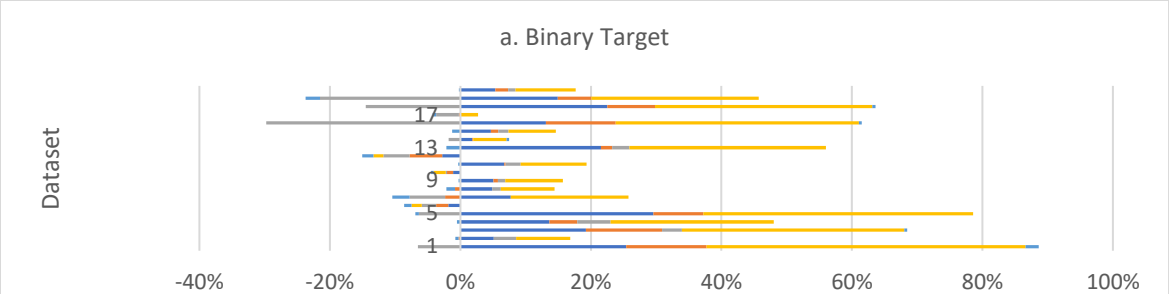
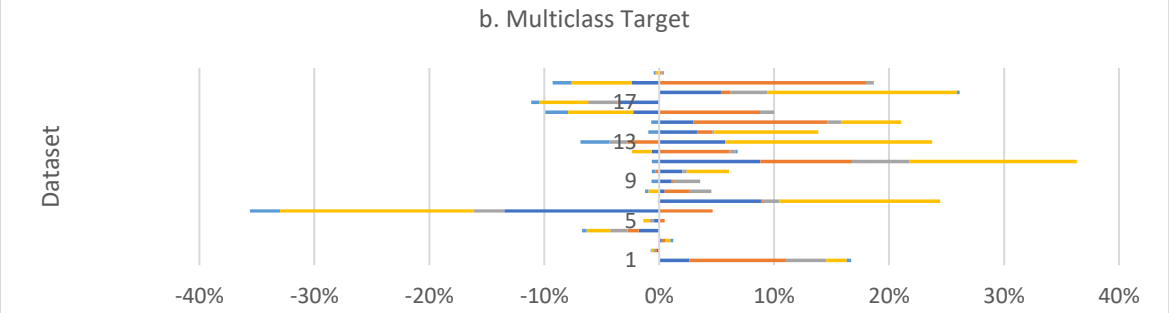



*Figure 1: Average prediction improvement after performing the data enhancement process.*

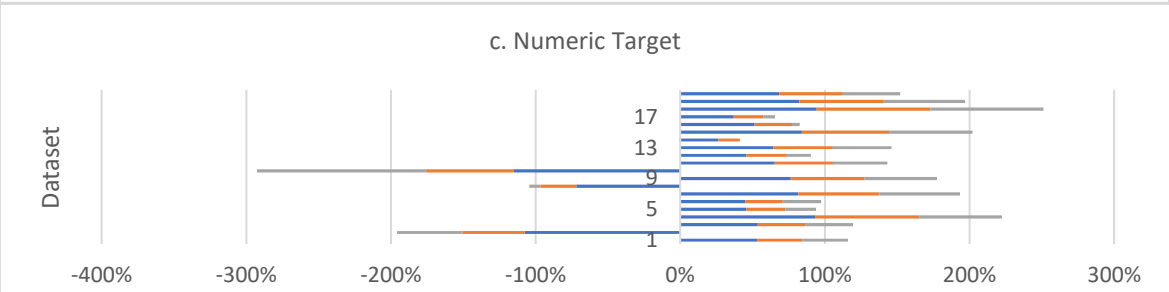
The detailed results for each dataset are presented in Figure 2.




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F1 Score | 25% | 5% | 19% | 14% | 30% | -2% | 8% | 5% | 5% | -1% | 7% | -3% | 22% | 2% | 5% | 13% | 0% | 23% | 15% | 5% |
| Balanced Accuracy | 12% | 0% | 12% | 4% | 8% | -2% | -2% | -1% | 1% | -1% | 0% | -5% | 2% | 0% | 1% | 11% | 0% | 7% | 5% | 2% |
| Precision | -6% | 3% | 3% | 5% | -6% | -2% | -5% | 1% | 1% | 0% | 2% | -4% | 3% | -2% | 2% | -30% | -3% | -15% | -22% | 1% |
| Recall | 49% | 8% | 34% | 25% | 41% | -2% | 18% | 8% | 9% | -2% | 10% | -2% | 30% | 5% | 7% | 37% | 3% | 33% | 26% | 9% |
| ROC AUC | 2% | -1% | 0% | -1% | -1% | -1% | -3% | -1% | 0% | 0% | 0% | -2% | -2% | 0% | -1% | 0% | -1% | 0% | -2% | 0% |


| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F1 Score | 3% | 0% | 0% | -2% | 0% | -13% | 9% | 0% | 1% | 2% | 9% | -1% | 6% | 3% | 3% | -2% | -3% | 5% | -2% | 0% |
| Balanced Accuracy | 8% | 0% | 0% | -1% | 0% | 5% | 0% | 2% | 0% | 0% | 8% | 6% | -3% | 1% | 12% | 9% | 0% | 1% | 18% | 0% |
| Precision | 3% | 0% | 0% | -1% | 0% | -3% | 1% | 2% | 2% | 0% | 5% | 1% | -2% | 0% | 1% | 1% | -3% | 3% | 1% | 0% |
| Recall | 2% | 0% | 0% | -2% | -1% | -17% | 14% | -1% | 0% | 4% | 15% | -2% | 18% | 9% | 5% | -6% | -4% | 17% | -5% | 0% |
| ROC AUC | 0% | 0% | 0% | 0% | 0% | -3% | 0% | 0% | -1% | 0% | -1% | 0% | -2% | -1% | -1% | -2% | -1% | 0% | -2% | 0% |




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MSE | 53% | -108% | 54% | 93% | 46% | 45% | 82% | -72% | 76% | -115% | 65% | 46% | 64% | 26% | 84% | 51% | 37% | 94% | 82% | 68% |
| RMSE | 31% | -43% | 33% | 72% | 27% | 26% | 56% | -25% | 51% | -61% | 40% | 28% | 41% | 15% | 60% | 26% | 21% | 79% | 58% | 44% |
| MAE | 32% | -45% | 33% | 57% | 21% | 27% | 56% | -8% | 50% | -117% | 38% | 17% | 41% | -1% | 58% | 5% | 8% | 78% | 56% | 40% |

*Figure 2: Detailed results for each dataset, demonstrating the impact of our data enhancement process*

## Conclusions

Our experiments demonstrate how Datomize's data enhancement process significantly improves machine learning model performance. By optimizing the training data, the prediction results dramatically improved when using the exact same models. Improvement was seen across the board, for most of the tasks, for each metric to varying degrees. For example, for 80% of the binary target tasks examined, there was an average increase of 12.6% in the F1 score; for 75% of the multiclass target tasks examined, there was a 5% increase in balanced accuracy; and for 85% of the numeric target tasks, there was an MSE reduction of 63% and an RMSE reduction of 42%.

The model-centric approach to AI now serves as the basis of many widely-used open-source models. The recent emergence of the data-centric approach to optimizing AI models has been driven by the ongoing need to improve model performance further and the search for new ways to accomplish this. Generative AI is rapidly rising in prominence as it moves to the public domain where it is taking center stage before a wider audience fascinated with its capabilities. However, for the last few years it has been explored in depth by industry and academic researchers, and generative AI now serves as a source of a wide range of content ranging from images and videos to program code. Given its demonstrated capabilities, it was only natural to consider leveraging generative AI's strengths to improve data quality. And from this coupling, Datomize emerged.

Our results highlight the important role that a generative AI approach can fill in a data-centric approach to improving model performance. Datomize’s data-centric generative AI-based approach transforms the data we collect into the data we want and need for improved performance.