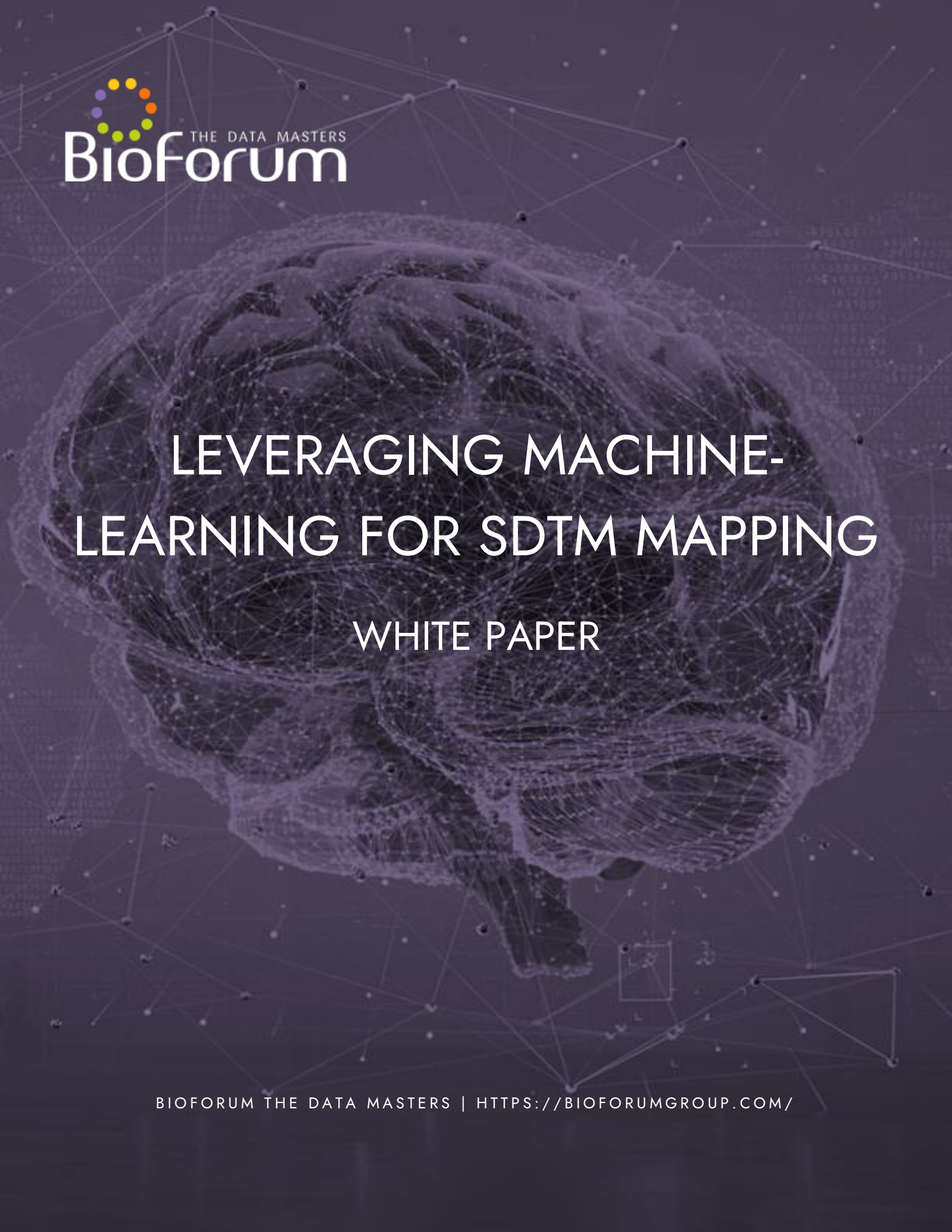


BioForum THE DATA MASTERS

LEVERAGING MACHINE-LEARNING FOR SDTM MAPPING

WHITE PAPER

EXECUTIVE SUMMARY

Data standards support the robust organization, analysis, and reporting of clinical trials and many regulatory authorities mandate the use of standardized data structures for clinical trial submissions.

The Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM) is a framework used to represent trial data in a standardized format for review. The SDTM structure has been adopted as the required data structure of many regulators.

Mapping raw clinical trial data to the SDTM framework can be tedious since raw data formats are diverse and may change during trial execution, resulting in repetitive conversion validation cycles. The initial mapping of source variables to SDTM domains and variables is a key step in the SDTM conversion process.

This paper introduces a machine-learning approach to generate SDTM domain mapping recommendations for domain and variable targets, discusses the accuracy of the underlying models, and presents refinement steps to improve the accuracy of model predictions.

JETConvert, Bioforum's next-generation SDTM conversion platform leverages the machine-learning solutions described in this paper for SDTM mapping.

TABLE OF CONTENTS

INTRODUCTION

BACKGROUND

Clinical trial data are diverse and influenced by collection methods, trial indication and research objectives. The complexity and amount of data collected during trials is increasing, driven by the wider use of, *inter alia*, adaptive trial designs, wearable devices, real world data, and an evolving landscape of guidance documents and therapeutic area user guides.

Many regulatory authorities, such as the Food and Drug Administration, require trial data to be submitted to them in a standardized format [1], namely, the Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM) [2]. However, the structure of collected data files is varied and data-collection system dependent. Contract research organizations, having a wide customer-base, typically receive data in dissimilar customer-specific formats adding complexity to the conversion of raw data to the SDTM framework [3].

Although the adoption of uniform data collection standards has reduced the effort required to produce a submission ready SDTM database, the transformation of raw data files to the SDTM standard remains a tedious manual process [4]. The mapping process is iterative per raw data variable: determine the target SDTM domains, identify target variables within the target domain and apply controlled terminology to the variable values. The current industry practice is to map raw variables to the SDTM standard in a specification and convert the raw clinical data, per the specification, using project-specific data transformation programs [5].

PROBLEM STATEMENT

Source-to-target SDTM mapping requires repetitive decision making. The manual conversion of raw data from a single clinical trial to the SDTM framework can consume weeks of work by experienced clinical programmers [6].

Is it possible to automate the transformation process, maintain the flexibility of SDTM standards, and capitalize on the experience of SDTM and trial experts?

OPPORTUNITY

Artificial intelligence, machine learning in particular, is well-established technology for automating repetitive decision making [7]. Can we leverage this opportunity to use machine learning to assist in the preparation of SDTM submission-ready packages?

SOLUTION

APPROACHING THE OPPORTUNITY

Bioforum saw several opportunities to use machine-learning as part of automating the preparation of SDTM submission-ready packages. This paper focuses on the application of machine-learning to determine the SDTM domains to which a raw variable will map and identify target SDTM variables within the selected domains.

Extract raw data features

Predict target domains & variables

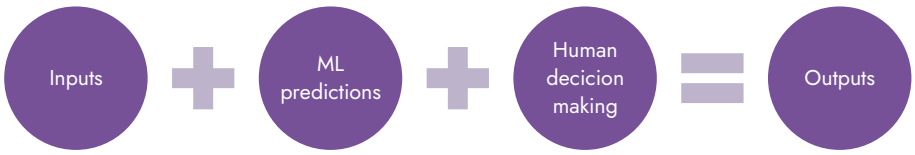
Accept or reject predictions

Output SDTM datasets & artifacts

The solution discussed in this paper placed the application of machine-learning within a broader workflow; using models to predict target SDTM domains and variables for evaluation by experienced staff.

The workflow included steps to:

- Extract raw data features from inputs: raw data from a variety of sources and associated trial documents (protocol, case report form (CRF), etc.).
- Use the extracted input features in machine-learning models to recommend the target SDTM domains and variables.
- Allow SDTM and trial experts to accept or reject the proposed mapping.
- Employ machine processing, rather than machine-learning, to output submission artifacts: SDTM domains, aCRF and Define-XML, etc.



The solution in this paper was extended for use in other parts of the SDTM submission-ready package preparation process, such as the recognition and mapping of variable values to SDTM Controlled Terminology [8].

HIGH-LEVEL SOLUTION

The machine learning approach used supervised learning algorithms to infer mapping recommendations. The algorithms were trained using pre-mapped trials from which raw variable features and associated SDTM mapping decisions were extracted. Multiple algorithms were trained and evaluated to select the best performing models per business case.

Thereafter, the selected models underwent refinement steps to improve the accuracy of the models to support SDTM mapping decisions.

TRAINING SET

CREATE A TRAINING SET

Preparing a training set is a foundational step in the implementation of machine learning. Supervised learning algorithms build models that generalize from existing (specific) samples to an entire universe. To effectively train an algorithm a reasonably large training set, that represents the universe the algorithm is expected to generalize to, is required. The training set should not be biased towards a specific region within the universe.

To build robust supervised learning models, equipped to handle a variety of clinical trials, the training set represented a range of therapeutic areas, research phases, trial sponsors, and collections systems. Moreover, the training set included legacy and ongoing trials.

The training set comprised clinical trials pre-mapped by the Bioforum Biometrics Team. Each raw variable was labeled with target SDTM domains and variables.

DIVERSE PRE-MAPPED TRIALS

- Therapeutic areas
- Research phases
- Trial sponsors
- Collection systems
- Legacy/ ongoing trials

FEATURE SOURCES

- Raw data file characteristics
- Raw variable metadata
- Summarized variable values
- Trial documents

EXTRACT DATA FEATURES

Hundreds of explaining features were extracted, from the following sources, to characterize the raw data:

- Raw data file characteristics, e.g., file name, label and the number of rows and columns in the file.
- Raw variable metadata, e.g., variable name and label.
- Summarized variable values, e.g., mean and standard deviation of the raw variable.
- Trial documents, e.g., CRF and the study protocol.

ADAPT THE TRAINING SET

The training set was modified to train the models to respond to common mapping scenarios. For example:

- Raw variables mapping to multiple targets were repeated in the training set by including a record for each target and repeating the raw variable features in each record.
- Variables, such as data-collection system variables, that do not get mapped to the SDTM framework, were filtered out of the set.
- To avoid the impact of rare occurrences when training the algorithm (e.g., overfitting to a rare domain) domains that appeared in two studies or less, such as DX, DI, and DT, were pooled. Similar adjustments were made on a variable-level.

MODELLING

MODELLING APPROACHES

Using supervised learning, we examined two modelling techniques. Firstly, single multi-class tasks which results in one model with a class for each domain or variable, depending on the application. Secondly, multiple binary classification algorithms which produce several models: in our case a distinct model for each domain or variable. The latter approach was found to be most suitable for our business cases.

BUILDING THE MODELS

Various algorithms were exposed to the training set, including logistic regression, Support Vector Machine with different kernels, naïve Bayes, XGBoost, artificial neural networks and random forests. The resulting models were then evaluated to determine the best model for each use case.

EVALUATING THE MODELS

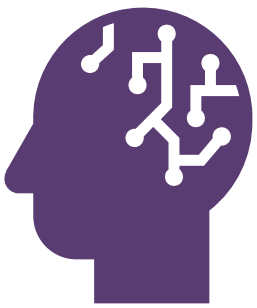
The performance of models was evaluated using cross validation methodology, i.e., "Leave-one-study-out" testing. The method iteratively selects a trial from the training set and the models provide predictions for the raw variables from the selected trial using the data from the remaining trials in the training set.

The testing results in a vector of probabilities for each raw variable where a probability represents the likelihood of the raw variable mapping to a target domain or variable that is present in the training set

As a starting point, to test model performance, we applied the simplest decision rule: Select the target domain or variable with the highest probability in the vector. Using this rule, we compared the predicted value (the target with the highest probability) to the expected value (the pre-mapped target from the training set) to select the models most compatible with our applications.

SELECTING THE MODELS

The random forest models yielded mapping predictions closest to the pre-mapped domain and variable labels in the training set, and in many cases, provided improved mapping recommendations.



MODEL PERFORMANCE

A confusion matrix is an effective way to present an overview of models' performance. The confusion matrix for the selected domain prediction model is depicted in Figure 1. The rows represent the pre-mapped domain targets, and the columns represent the domain targets associated with the highest probability in the likelihood vector. Using the naïve decision rule, mentioned in the previous section, raw variables could only be mapped to a single domain.

Frequencies on the main diagonal of the matrix are higher than in the other cells, indicating that random forest classifiers are effective at predicting SDTM domain targets for raw variables. For domain-level models the overall accuracy is 71.5%, based on 41 pre-mapped clinical trials. The matrix is also useful to identify the domains containing variables being erroneously mapped to a different domain, for example, raw variables to be mapped to PR are most often mistakenly mapped to LB.

*Accuracy is a measure of the model predicting correctly versus all the predictions that the model is making. For **domain-level models** the overall accuracy is **71.5%** and, for the **variable-level models**, **83.8%***

A similar confusion matrix, based on 61 pre-mapped clinical trials, was created to evaluate the variable-level models. The size of the variable-level matrix restricts its presentation in this paper; however, the overall accuracy of the models was 83.8%.

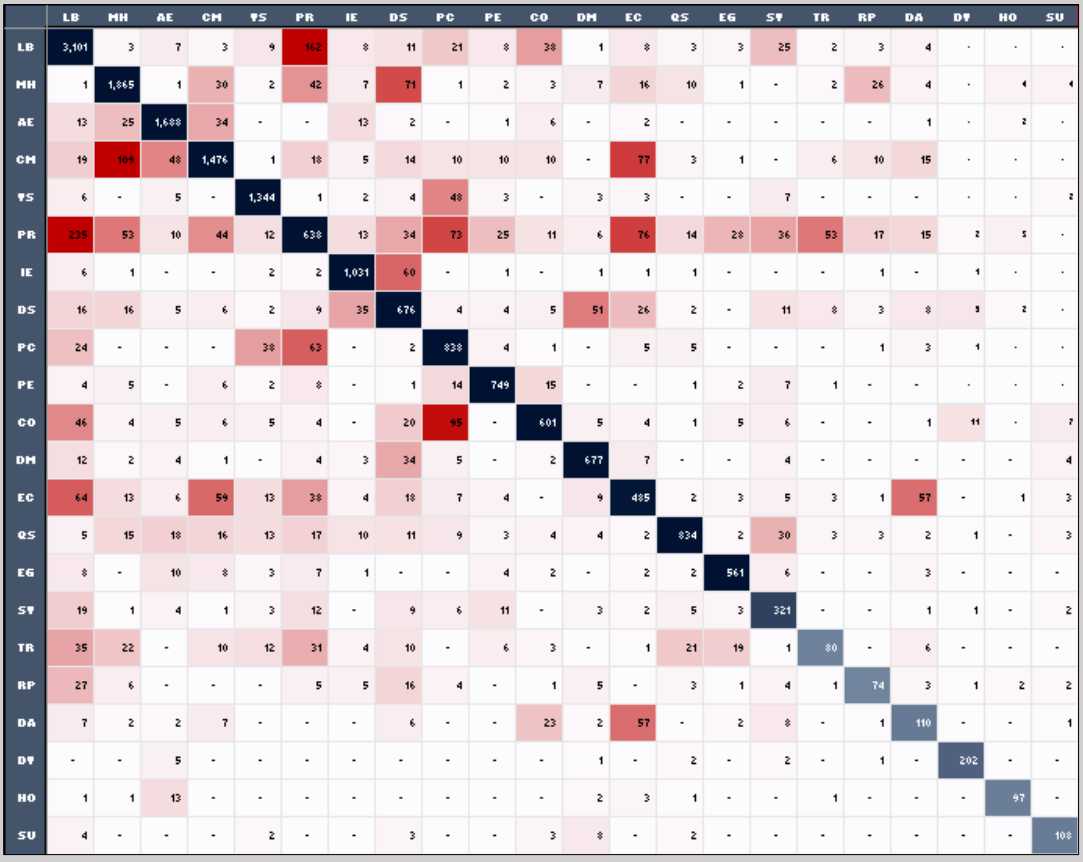


Figure 1: Confusion matrix for domain-level models using the simple "maximal probability" rule, non-pooled domains

IMPROVING PERFORMANCE

The domain and variable models were deployed using Bioforum's innovative workflow based SDTM conversion tool. However, user confidence in the tool's domain[1] and variable[2] target accuracy was not high enough to motivate users to use the new tool. To improve the models' overall accuracy a series of business implementation steps were applied to the domain and variable model.

DOMAIN MODEL IMPROVEMENTS

STEP 1 – REMOVE PREDICTIONS WITH LOW CONFIDENCE

STEP 2 – PROVIDE THE TOP 3 MOST LIKELY PREDICTIONS

STEP 3 – DYNAMIC PREDICTIONS BASED ON CUMULATIVE CONFIDENCE THRESHOLDS

STEP 1

User feedback indicated a preference for a tool that recommends domain targets with a high likelihood of being correct, omitting predictions associated with low confidence. In the computing industry, this is known as increasing the model's precision at the price of lower model recall.

Thus, the first refinement step was to remove model predictions if the likelihood associated with the prediction is lower than a confidence threshold. Figure 2 illustrates the trade-off between recall and precision by confidence threshold.

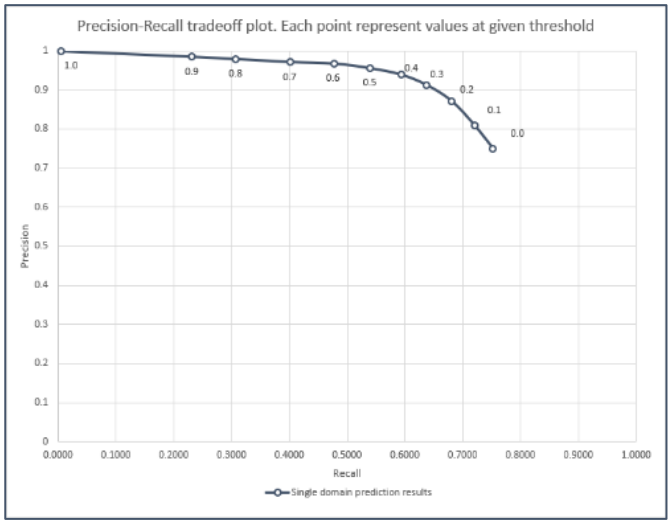


Figure 2: Precision-Recall trade-off for a single option result model

[1] Overall domain-level models' accuracy of 71.5%, based on a training set containing 41 clinical trials

[2] Overall variable-level models' accuracy of 83.8%, based on a training set containing 61 clinical trials

STEP 1 (continued)

At a high confidence threshold, few predictions reach the threshold, resulting in a low recall. However, the few predictions that reach the threshold have a higher chance of matching the pre-mapped variables (higher precision). Conversely, at a low threshold, multiple predictions exceed the threshold, yet fewer predictions match the pre-mapped variables.

A confidence threshold of 0.30 represented a reasonable position on the precision-recall curve. Removing domain predictions associated with a confidence level of 0.30 or less yielded a precision of 91.4%, with a recall of 63.7%.

The confidence threshold model configuration significantly increased user trust in the domain-level models. For some of the trials, the configuration led to nearly error-free recommendations. Overall, the refined approach resulted in 69.7% of the raw variables receiving a mapping recommendation. The remaining variables did not have any domain predictions associated with a confidence level exceeding the threshold.

STEP 2

Domain model prediction errors detected after Step 1 were analyzed to understand opportunities to further improve model performance. It was observed that if the prediction associated with the highest confidence did not match the pre-mapped target, the sought-after recommendation was often in the second or third position. Users requested that the tool provide a list containing the three most likely target domains with functionality allowing users to select the preferred domain from the list.

The confidence threshold model configuration was adapted to provide the three most likely domain recommendations. The percentage of cases in which the sought-after domain recommendation, or at least one sought-after recommendation, was included in the three recommendations improved to 96.6% with a recall of 67.3% at the threshold of 0.3. Figure 3 displays the overall precision-recall graph.

The impact of Step 2 was seen in user-satisfaction. Users were able to easily select the correct domain when the most likely target was not suitable. However, this approach required users to constantly evaluate three options, even when the acceptable target was evident.

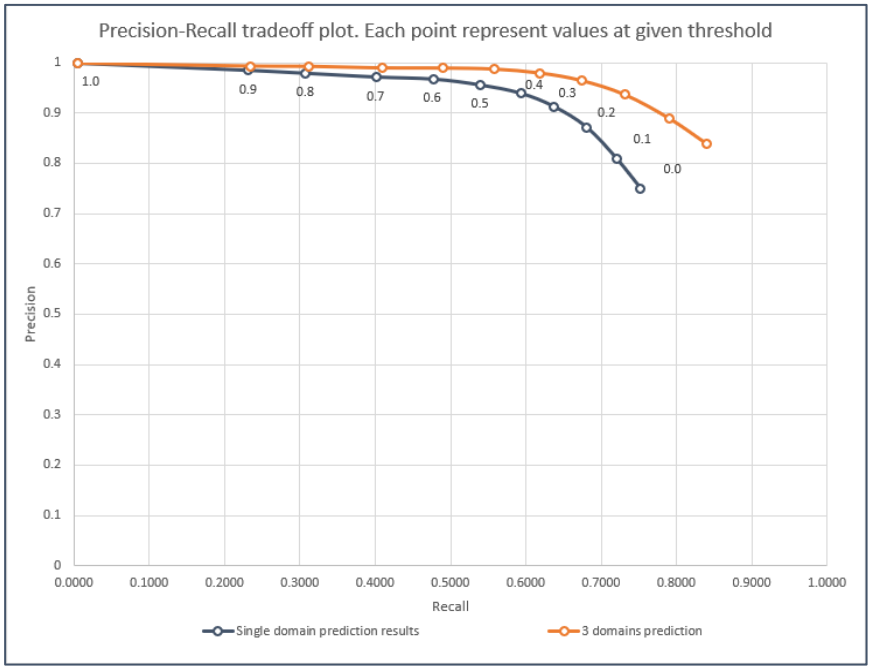
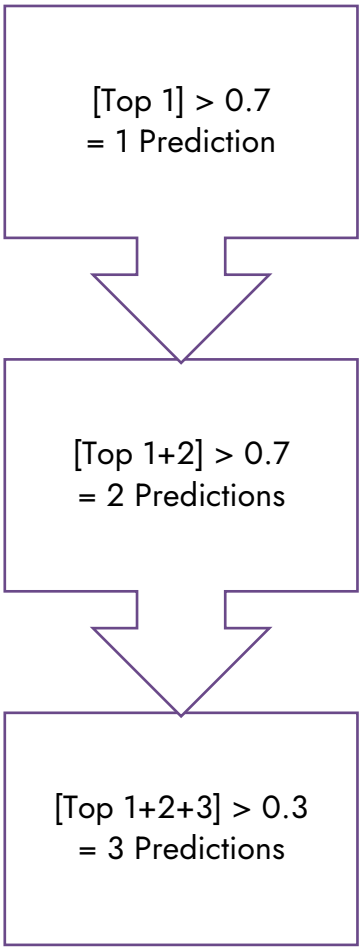


Figure 3 Precision-Recall trade-off when providing 3 domain predictions



STEP 3

Step 3 was investigated as a method to maintain the benefit of removing predictions associated with low confidence, provide a single prediction when the suitable target was evident, and provide alternatives when the target was not apparent.

This refinement step adapted the number of recommendations provided to a user based on cumulative confidence thresholds.

- If the confidence in the most likely target is high enough ➔ present a single target to the user,
- If not, consider the combined confidence of the top 2 most likely targets, if the confidence is high enough-➔ present two targets,
- If not, consider the combined confidence of the top 3 most likely targets, if the confidence is high enough-➔ present three targets (as implemented in Step 2)

The precision-recall curve for this method is displayed in Figure 4, for the specific application of thresholds of 0.70, 0.70 and 0.30. This model configuration provided an acceptable tradeoff by maintaining high precision, yet still provided a single recommendation for most variables (about 60% of predictions).

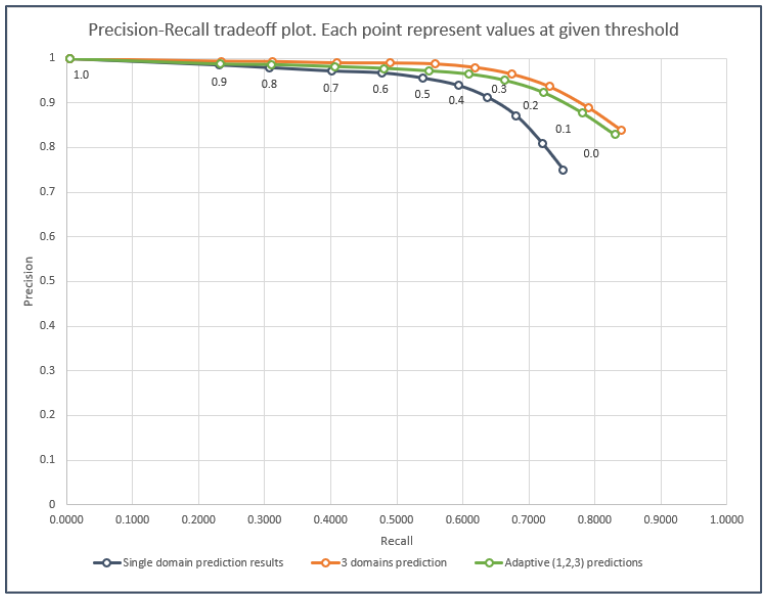


Figure 4: Precision-Recall trade-off when providing an adaptive cumulative threshold approach

VARIABLE MODEL IMPROVEMENTS

As mentioned in the Modelling section, random forest models yielded mapping predictions closest to the pre-mapped variable labels in the training set. The overall accuracy of these models, when implementing the simple decision rule of selecting the target variable with the highest likelihood for each raw variable, was 83.8%[3]. A refinement process, like the one applied to the domain-level models, was employed to improve the accuracy of the variable-level models.

STEP 1

Provide predictions with a likelihood above a static threshhold, up to a maximum of 3 targets

STEP 2

Provide a dynamic number of predictions based on cumulative confidence threshholds

STEP 1

When mapping data on a variable level, users prefer being presented with multiple mapping recommendations and the functionality to select the correct target. Thus, the models were configured to provide users with targets associated with a likelihood above a static threshold, up to a maximum of three targets.

The precision-recall curve (Figure 5) shows that:

- At a threshold of 0.05, the maximum number of three targets is being provided for all variables, thus recall is 90%, but precision is low at 40%.
- Once again, a confidence threshold of 0.3 provides an acceptable balance between precision (86.9%) and recall (78.6%).

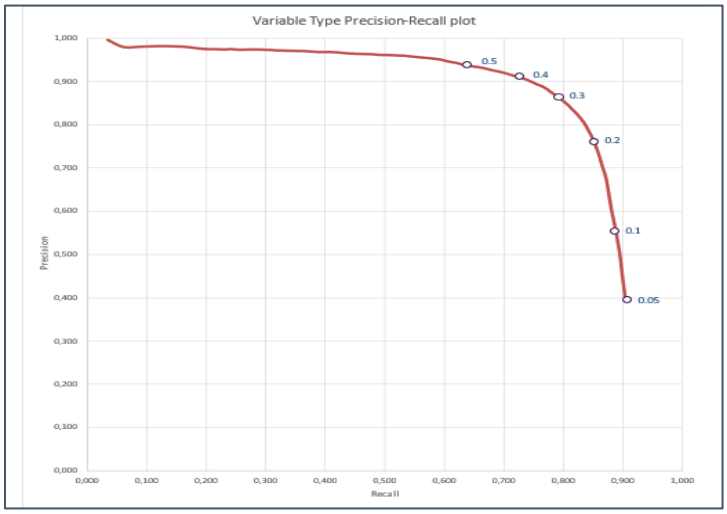


Figure 5: Precision-Recall trade-off when providing up to three variable predictions above a static threshold

STEP 2

Investigating the method whereby a dynamic number of predictions is provided per variable, using the step-down approach that was introduced in Step 3 for the domain models, did not result in improvements in precision or recall. Thus, Step 1, presented above, provided the best results with respect to the refinement of the variable-level models.

[3] Based on a training set containing 61 pre-mapped trials

CONCLUSION

This white paper discusses the application of machine-learning to generate SDTM domain and variable mapping recommendations. The approach reduces repetitive decision making and manual processing associated with converting raw data to the CDISC SDTM framework.

Supervised learning algorithms were trained using pre-mapped trials from which raw variable features and SDTM domain and variable targets were extracted. Given raw variable features, the models assign a likelihood of mapping to each SDTM domain and variable in the training sets. The simplest application of the models, selecting the most probable targets, was 75.1%[4] accurate for domain predictions and 83.9%[5] accurate for variable predictions, respectively.

A series of business implementation steps were effective in improving the overall accuracy of model predictions. Applying individual and cumulative confidence thresholds in combination with decision rules to determine the number of predicted values displayed to users, improved the prediction accuracy to 96.6% and 86.9% for the domain and variable models, respectively.

JETConvert, Bioforum's SDTM automation solution, uses machine-learning to produce source-to-target domain mapping predictions for evaluation by trial experts. **JET**Convert integrates a machine learning approach with an innovative workflow-based system to produce a submission-ready SDTM package.

[4] Based on a training set containing 41 pre-mapped clinical trials

[5] Based on a training set containing 61 pre-mapped clinical trials

REFERENCES

[1] U.S. Food & Drug Administration, "Data Standards Catalog," Rockville, 2022.

[2] CDISC, "Study Data Tabulation Model," 29 November 2021. [Online]. Available: https://www.cdisc.org/standards/foundational/sdtm. [Accessed 25 November 2022].

[3] S. Vijendra,et al, "The Elusive Goal of Automation in SDTM Mapping: A CRO Perspective," in PHUSE Connect, Orlando, 2020.

[4] PHUSE, "Industry Experiences Submitting Standardized Study Data to Regulatory Authorities," 15 September 2020. [Online]. Available: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Optimizing+the+Use+of+Data+Standards/Industry+Experiences+Submitting+Standardised+Study+Data+to+Regulatory+Authorities.pdf. [Accessed 25 November 2022].

[5] J. Fulton, "Ensuring Consistency Across CDISC Dataset Programming Processes," in PharmaSUG, Virtual Proceedings, 2020.

[6] J.E. Stuelpner, et al, "Data Transformation: Best Practices for When to Transform Your Data," in PharmaSUG, Virtual Proceedings, 2020.

[7] S. Brown, "Machine learning, explained," MIT, Sloan School of Management, 21 April 2021. [Online]. Available: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained. [Accessed 13 December 2022].

[8] CDISC, "Controlled Terminology," 30 September 2022. [Online]. Available: https://www.cdisc.org/standards/terminology/controlled-terminology. [Accessed 14 December 2022].