phuse·share
connect·share·advance

Single Day Event

A Guided User Experience for Automated Mapping

Indrasena Reddy Purini
Director, Elluminate Services.
eClinical Solutions

Key Challenges in Clinical Trial Data Submission

- Data Quality & Integrity
- Regulatory Compliance
- Timeliness & Efficiency
- Interoperability & Standardization
- Technology Limitations

Benefits of Automation

- Improved Data Quality & Integrity
- Ensuring Regulatory Compliance
- Enhanced Cycle Times
- Accelerated Data Standardization
- Simplifying the Complex
- Leveraging Technology Innovation

Data Mapping Steps

Specification

- Describes the documentation of how mapping is to be performed. This step is usually performed by a programmer or subject matter expert who understands both the source and the target data standards.

Execution

- Carrying out of the mapping defined in the specification, typically by a programmer who translates the mapping instruction into code.

Approach

To Automapping

Data Classification

- Use ML techniques to automatically classify study metadata to a standard model. Store the classified metadata so it can be used by other platform components.

Spec Creation

- From classified study metadata, and given a selected target standard, automatically generate a mapping specification document.

Spec Transformation

- With a client-provided specification document, use NLP and other techniques to interpret human readable mapping notes and translate the document to the Automation-Ready Mapping Specification template.

Mapping Code Generation

- Take an Automation-Ready Mapping Specification document and from it generate mapping code in SAS, R, or Python.

Submission Document Generation

- With a Mapping Specification and other necessary metadata, generate other documents required for submission packages

Solution Workflow

To Build and Deploy Machine Learning Models

1. Corpus Collection
2. Ground Truth Labeling
3. Pre-processing and Feature Engineering
4. Model Selection and Development
5. Model Evaluation and Deployment

Corpus Data Collection

Components and Methods

- Standards
  - SDTM IG v3.3/3.4
  - Control Terminology
- Knowledge Base
  - Key Master
  - Domain Master
- Source Data
  - Metadata Collection (250+ studies)
- Synthetic Data
  - Algorithm Generated Data

Ground Truth Labeling

Annotation Details

- Total observations : 845990
- Columns: 15
- variables (Excluding SEQ) :
  - *STUDYNAME, DOMAIN, VARIABLE, LABEL, ORIGIN, KEY, ROLE, TYPE, LENGTH, FORMAT, CODELIST, PAGENO, CDISC_NOTES, CORE, MAPPING_NOTES*
- DOMAINS: 56
- No. of Resource who has contributed for this task : 2

Exploratory Data Analysis

Key highlights

Data Preparation :

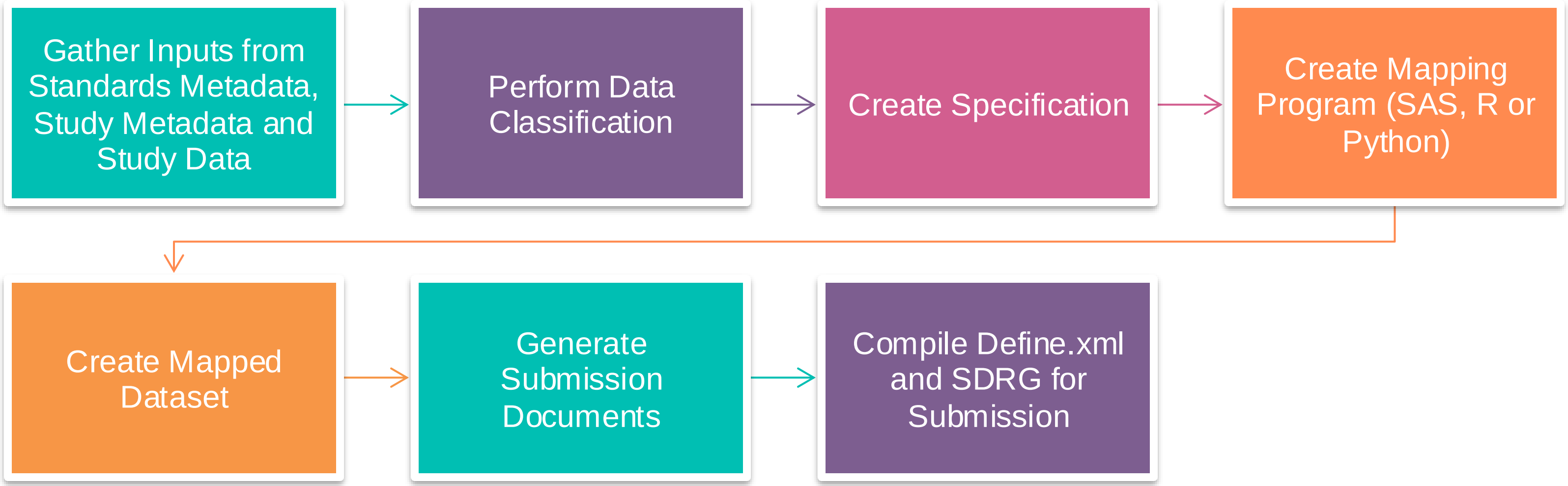
Collected 250+ studied meta data and performed EDA and verify any outliers as per EDA process (univariate, bivariate, boxplot, heatmap etc..)

Missing Value Treatment :

For most of the missing values, we imputed based on the MODE (for categorical) / MEAN (for numerical values)

Automated Mapping

Process & Initiatives



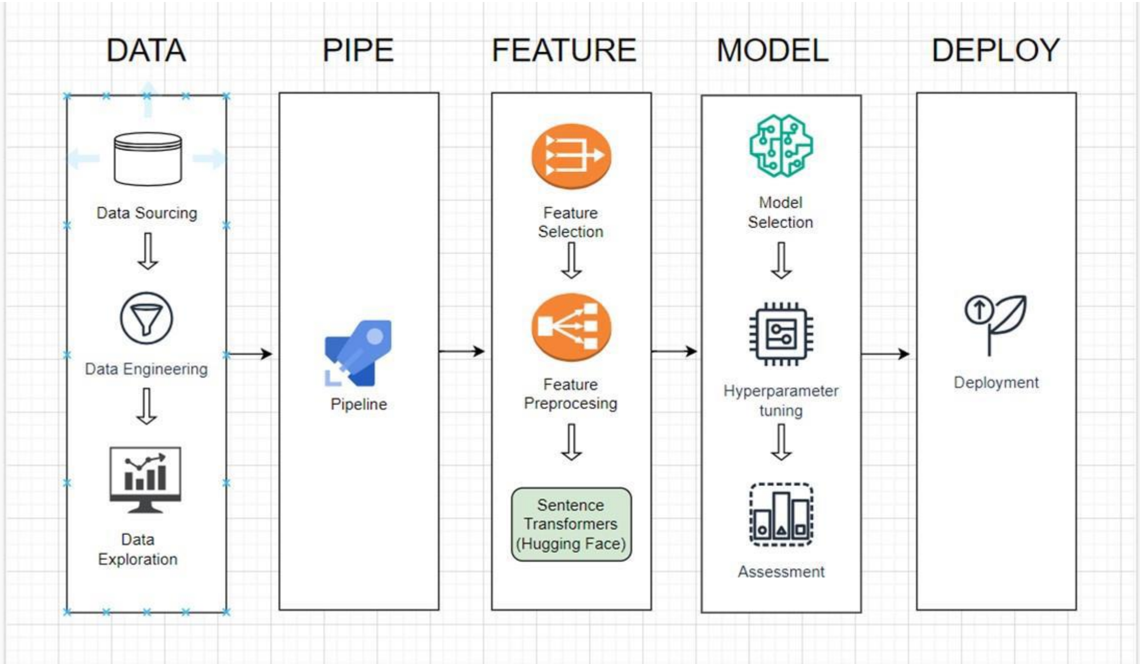
Data Classification

Mapping Types

- Direct
- Coalesce
- Concatenation
- Substring
- Condition
- DateDiff
- Datetime
- Decode
- Imputation
- MaxOf
- MinOf
- Scan
- Sequence
- TextInsert
- Aggregation
- Transpose

Designing & Deploying

Ensemble Models



Benefits of Ensemble Models

For Automapping

Improved Accuracy	Robustness	Reduced Overfitting	Enhanced Predictive Power
Flexibility	Scalability	Error Reduction	Adaptability

Modeling

Approach

- Ensemble with:
  - Majority Voting Classifier with Sentence Transformation
- Final prediction is determined by the class label that receives the most votes from the individual models.
- In case of a tie, the algorithm can employ various tie-breaking strategies, such as selecting the first or last prediction.
- Some of the optimal parameters which are used for classification :
  - learning_rate ,
  - n_estimators,
  - max_features,
  - max_depth,
  - min_samples_split,
  - random_state etc..

Selected Machine Learning Models (1/2)

For the Ensembles

Logistic Regression (LR):

- Description: A statistical model used for binary classification tasks.
- It predicts the probability of a binary outcome (e.g., yes/no, true/false) based on one or more predictor variables.
- Use Case: Commonly used in medical diagnosis, credit scoring, and marketing.

Random Forest (RF):

- Description: An ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees.
- Use Case: Used in various applications like fraud detection, disease prediction, and stock market analysis.

Support Vector Machine (SVM):

- Description: A supervised learning model that finds the optimal hyperplane which maximizes the margin between different classes in the feature space.
- Use Case: Effective in high-dimensional spaces and used in text classification, image recognition, and bioinformatics.

K-nearest Neighbor (KNN):

- Description: A simple, instance-based learning algorithm that classifies a data point based on how its neighbors are classified.
- It uses the majority vote of its k-nearest neighbors.
- Use Case: Used in recommendation systems, pattern recognition, and intrusion detection.

Selected Machine Learning Models (2/2)

For the Ensembles

Multi-Layer Perceptron (MLP):

- Description: A type of artificial neural network with one or more hidden layers between the input and output layers.
- It uses backpropagation for training.
- Use Case: Applied in speech recognition, image classification, and machine translation.

Extra Trees (ET):

- Description: An ensemble learning method similar to Random Forest but with more randomness.
- It builds multiple trees using random splits of all observations and features.
- Use Case: Used in regression and classification tasks where high variance is a concern.

Decision Tree (DT):

- Description: A tree-like model used for decision making.
- It splits the data into subsets based on the value of input features, creating branches until a decision is made.
- Use Case: Used in customer segmentation, risk analysis, and medical diagnosis.

Bernoulli-Naïve Bayes:

- Description: A variant of the Naïve Bayes classifier for binary/boolean features.
- It assumes that the presence of a particular feature is independent of the presence of any other feature.
- Use Case: Commonly used in text classification tasks like spam detection and sentiment analysis.

Outcome & Selection

Based on F1 scores and performance

	Ensemble A	Ensemble B	Ensemble C	Ensemble D	Ensemble E
Description and estimators	8 Models (LR, RF, SVM, KNN, MLP, ET, DT and BNB)	5 Models (LR, RF, SVM, KNN, MLP)	3 Models (RF, SVM, MLP)	5 Models (LR, RF, SVM, KNN, MLP)	4 Models (LR, RF, SVM, MLP)
Features	5	5	5	16	5
Accuracy	0.862	0.887	0.889	0.875	88.86
F1 Score	0.606	0.662	0.655	0.626	0.787
Execution Time for (50 sample size)	30 to 40 mins	8 mins	5 mins	15 mins	Below 15 secs
Precision	0.638	0.679	0.623	0.680	0.807
Recall	0.603	0.668	0.655	0.660	0.783
Comments	Since this model takes time to load and execution, this has to be eliminated.	This model execution time is 8 mins , score is decent and 5 set of estimators.	This model execution time is 5 mins and score is decent. However it has only 3 estimators.	This model execution time is 15 mins , score is below expectations. So this has to be retired.	Fine tuning the code to execute faster
Next Steps	Eliminated	1. Adding more corpus to improve the scores. 2. Finetune Hyper Parameters 3. Adding more features which are appropriate to model	Eliminated	Eliminated	1. Finetune of Hyper parameters 2. Adding more corpus for including control values

Conclusion

The Use of Ensembles for Automapping

- Automation of spec generation and mapping is a significant advancement
- Ensemble of ML models provides a good result
- Addresses many challenges of clinical trial data standardization
- Substantial reduction in cycle times, effort and decrease in error rates
- Need to embrace digital transformation
- Promising step towards efficient, accurate and reliable data submissions
- Faster treatments to market

Thank you

Indrasena Reddy Purini
Director, Elluminate Services
eClinical Solutions
ipurini@eclinicalsol.com

The Global Healthcare
Data Science Community

Q&A

The Global Healthcare
Data Science Community

Contact Channels
UK +44 1843 609600
office@phuse.global
phuse.global

Social Media
@phusetwitta
/phusebook
/phusetube
/company/phuse