WHITEPAPER

AI-Enabled Data Quality

In this paper we provide an overview of Artificial Intelligence (AI) and Machine Learning (ML) and their application to Data Quality. We highlight how tools in the Datactics platform can be used for key data preparation tasks including cleansing, feature engineering and dataset labelling for input into ML models.

A real-world application of how ML can be used as an aid to improve consistency around manual processes is presented through an Entity Resolution Use Case. In this Use Case we show how using ML reduced manual intervention tasks by 45% and improved data consistency within the process.

Published on 10th June 2020

Contents

1. The Role of Data Quality

Having good quality, reliable and complete data provides businesses with a strong foundation to undertake tasks such as decision making and knowledge to strengthen their competitive position. It is estimated that poor data quality can cost an institution on average $15 million annually [1]. A Forrester study estimated that only 12% percent of companies are applying data-driven intelligence to guide their company strategy and inform their business objectives[2]. As we continue to move into the era of real-time analytics and Artificial Intelligence (AI) and Machine Learning (ML) the role of quality data will continue to grow. As highlighted by Forrester[3] for companies to remain competitive, they must have in place flexible data management practices underpinned by quality data. This will provide data-driven insights to aid strategic business questions.

AI/ML are being used for predictive tasks from fraud detection through to medical analytics. These techniques can also be used to improve data quality when applied to tasks such as data accuracy, consistency, and completeness of data along with the data management process itself.

In this paper we will provide an overview of the AI/ML process and how Datactics tools can be applied in cleansing, deduplication, feature engineering and dataset labelling for input into ML models. We highlight a practical application of ML through an Entity Resolution Use Case which addresses inconstancies around manual tasks in this process.

2. An Overview of Artificial Intelligence and Machine Learning

There are many statistics around AI/ML and its anticipated impact on society and industry. For example, Gartner predicts the business value created by AI will reach $3.9T in 2022 and Forbes has reported that AI/ML will have the potential to create an additional $2.6T in value by 2020 in Marketing and Sales, and up to $2T in manufacturing and supply chain planning4 [4]. The AI/ML domain has certainly come a long way from the question posed by Alan Turing in the 1950s of whether machines can think. Often the term AI is used interchangeable with the term ML, however, there are differences between the two as detailed below.

> ***"Andrew Ng refers to AI as automation on steroids"***

Artificial Intelligence

The term AI was first coined by John McCarthy in 1956 when he invited a group of researchers from a variety of disciplines including language simulation, neuron nets, complexity theory and more to a summer workshop to discuss what would ultimately become the field of AI. Although there is no definitive definition of AI, it is broadly speaking the capability of a computer to perform cognitive tasks which are usually performed by a human. These include tasks such as decision making, visual perception and speech recognition. AI can be classified into **general** and **narrow** AI.

- **General AI**
  In general AI, the goal is to develop AI to the point where a computer's intellectual capability is equal to the cognitive capability of a human. Such as a general AI system which can function in

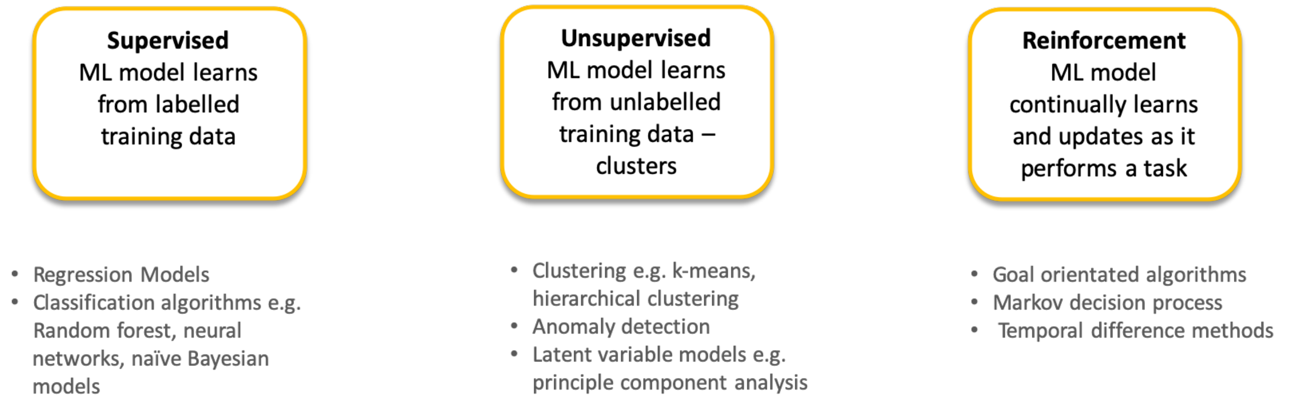
any scenario / sector without explicitly acting on rules imposed on it. At present, we are not at the stage of general AI. When general AI will be available is inconclusive with estimates in the next 50 years plus, to the argument that General AI may not be feasible.

- ❖ **Narrow AI**
  The common use of the term AI falls into the classification of narrow AI. A narrow AI system is applied to and trained to perform a specific task. Examples include Amazon Alexia, the personal assistant which uses natural language processing and natural language generation to understand and communicate. Within the medical sector we see the use of narrow AI applied to the prediction of disease and recommendation of clinical treatments based on the analysis of patient data.

Machine Learning

ML is a subset of AI. These are the models and statistical models such as neural networks, random forests, k-nearest neighbour, Markov chain models and Bayesian networks. These mathematical models learn from large volumes of data, identify patterns within this data and use this knowledge to make prediction on new unseen cases.

There are different types of ML and these can be broadly separated into supervised, unsupervised and reinforcement learning, these are summarised in Figure 1 below. Supervised learning is where a ML model learns from labelled training data, for example, predicting if an email is spam or not. When you have no labelled data, you can use unsupervised learning to identify clusters or patterns within your data such as grouping customers by what they purchase. Then you have reinforcement learning where an ML model learns as it performs a task for example, learning how to play the game AlphaGo as highlighted in the work by Deepmind[5].



Supervised
ML model learns from labelled training data

- Regression Models
- Classification algorithms e.g. Random forest, neural networks, naïve Bayesian models

Unsupervised
ML model learns from unlabelled training data – clusters

- Clustering e.g. k-means, hierarchical clustering
- Anomaly detection
- Latent variable models e.g. principle component analysis

Reinforcement
ML model continually learns and updates as it performs a task

- Goal orientated algorithms
- Markov decision process
- Temporal difference methods

Figure 1 Summary of the different type of ML approaches including models

In the financial sector we can see the diverse application of ML technologies including the profiling of behaviour in fraud detection, the use of NLP on unstructured data to extract information to enrich the Know-Your-Customer (KYC) onboarding process, through to the use of chat bots within customer service to automatically address queries from customers.

There are two key challenges in ML namely data quality and ensuring there is sufficient training data for an ML algorithm to learn from as detailed below.

1. **Data Quality:** This is often a time-consuming but important process performing pre-processing tasks such as deduplication, cleansing and labelling. For a machine learning engineer, undertaking ML research, an ad-hoc, one-time approach to data prep may be suffice. However, in production, ML pipelines and processes need to be reproducible. This will often involve the handling and managing of constantly changing data which requires the implementation of continuous data quality to constantly monitor and fix data. Ideally, there would be a data-prep pipelines to provide a standardised, transparent and reproduceable approach to cleansing and generating datasets for ML models. These data quality pipelines could then be saved and re-used in other ML projects.
2. **Training Data:** for supervised ML algorithm to make predictions, it requires several "examples" to firstly learn from. This learning involves extracting patterns from the training data and then using this information to make predictions on new unseen cases. The training dataset needs to be sufficient in size to identify patterns, avoid overfitting and contain accurate data which is representative of the ML problem.

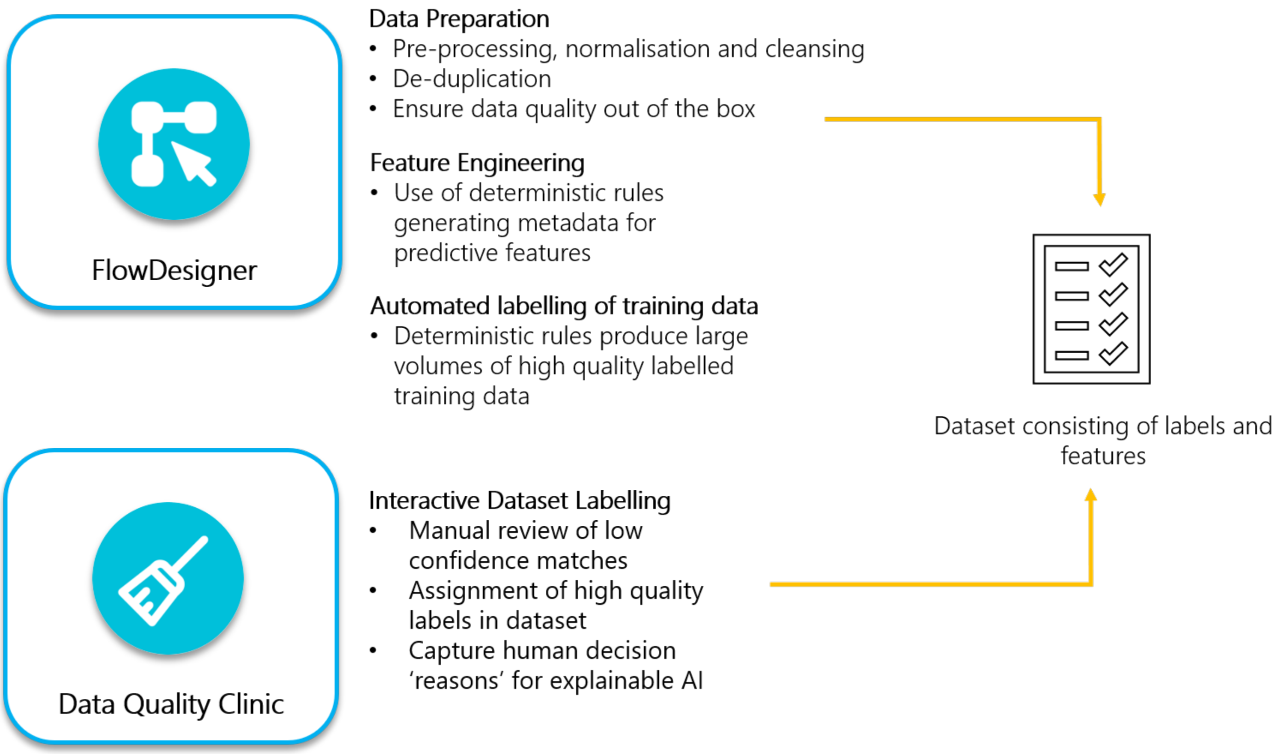
3. Datactics Tools for Dataset Cleansing, Feature Engineering and Labelling

ML models take as input datasets from which they extract patterns, insights and learnings, building knowledge from which predictions can be made. For example, analysing customer behaviour and using this knowledge to recommend products. These ML models require datasets to be correctly constructed, transformed into the appropriate structure and consisting of good quality, representative data of the prediction problem they are applied to.

> ***"It is a commonly held maxim in predictive analytics that 60%–90% of the project time will be consumed by data preparation tasks."***

Data preparation is an important step in the ML process, taking care in this step with cleansing and curating your data and addressing issues such as formatting, missing values will have an impact on the quality of predictions from the ML algorithm. It is however a time-consuming step. As expanded upon in the "Data Preparation Cookbook"[6], data preparation is more than just data cleansing. It involves tasks including deduplication, reformatting, standardisation and normalisation, noise reduction, ensuring data is accurate, non-biased and anonymised.

The ***Datactics Data Quality platform*** can perform such data preparation operations at scale by providing ready access to tools and processes that can ensure a base level of quality and identify anomalies in data that may skew ML models. Furthermore the same data quality tools can assist data scientists to generate metadata that can be used to train machine learning models – such 'Feature Engineering' can be of real value when the data set is largely textual as it can generate numerical indicators which are more readily consumed by ML models. A summary of these processes is presented below and summarised in Figure 2.



Data Preparation

- Pre-processing, normalisation and cleansing
- De-duplication
- Ensure data quality out of the box

Feature Engineering

- Use of deterministic rules generating metadata for predictive features

Automated labelling of training data

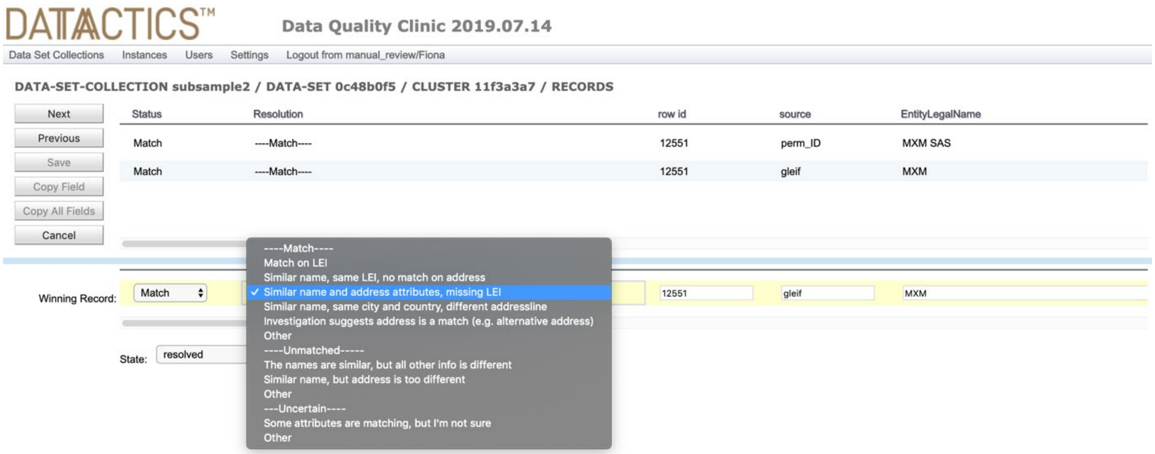
- Deterministic rules produce large volumes of high quality labelled training data

Interactive Dataset Labelling

- Manual review of low confidence matches
- Assignment of high quality labels in dataset
- Capture human decision 'reasons' for explainable AI

Figure 2. Summary of the Datactics platform tools used to prepare, engineer and label datasets for input into ML models

- **Data Preparation**
  Use of FlowDesigner drag and drop user-friendly interface to perform data quality operations such as cleansing, parsing, validating, formatting, deduplication, and enrichment by matching disparate data sets.
- **Feature Engineering**
  A feature is a measurable property of an object that you are making a prediction on. For example, the measurement of blood pressure to predict the risk of heart disease. The Datactics platform can be used to create predictive features for a dataset using deterministic match rules and the underlying match meta-data.
- **Dataset Labelling**
  For an ML model to learn, it is presented with representative examples of the task it is aiming to predict. For instance, what a high confidence match looks like and what a low confidence match looks like. The model needs labelled examples to undertake this task. Providing quality labelled data can be a time consuming and manual process. Often there is manual intervention requiring human input to labelling process. Within FlowDesigner, high confidence matches using deterministic rules can automatically provide large volumes of high-quality labelled data instances. Furthermore, the Data Quality Clinic provides an interactive dataset labelling interface where reviewers can provide high quality labels to the edge cases within their dataset along with capturing their decision rationale to enable **explainable AI** as presented in Figure 3.



DATACTICS™
Data Set Collections
Instances
Users
Settings
Logout from manual_review/Fiona
Next
Previous
Save
Copy Field
Copy All Fields
Cancel
----Match----
Match on LEI
Similar name, same LEI, no match on address
✓ Similar name and address attributes, missing LEI
Similar name, same city and country, different addressline
Investigation suggests address is a match (e.g. alternative address)
Other
----Unmatched-----
The names are similar, but all other info is different
Similar name, but address is too different
Other
---Uncertain----
Some attributes are matching, but I'm not sure
Other

Data Quality Clinic 2019.07.14

DATA-SET-COLLECTION subsample2 / DATA-SET 0c48b0f5 / CLUSTER 11f3a3a7 / RECORDS

Status	Resolution	row id	source	EntityLegalName
Match	----Match----	12551	perm_ID	MXM SAS
Match	----Match----	12551	gleif	MXM

Winning Record: Match | Similar name and address attributes, missing LEI | 12551 | gleif | MXM

State: resolved

Figure 3. The Data Quality Clinic interface where datasets can be labelled through the review process and decision rationale captured.

In the following section of the paper we outline a practical application of the Datactics platform tools for constructing and labelling a dataset for use in a real-world Entity Resolution (ER) Use Case. This Use Case provides an overview of research performed in measuring the application of ML to address inefficiencies and inconsistencies around manual tasks in the ER process.

4. AI-Assisted Entity Resolution Use Case

It is essential that financial institutions know their customers and have verified their identities so that they can comply with regulations, fraud detection and risk modelling. An important part of the KYC/AML process is ER which is the task of identifying and resolving entities from multiple data sources. This is often a manual and time-consuming task which has an impact on regulation compliance and customer experience. A report by Refinitiv[7] has highlighted that current onboarding processes can cost roughly around $28.5 million annually for an institution. Three main challenges in ER revolve around data quality, operational efficiencies, and customer experience. Interestingly, research by Forrester found that clients can expect a 20% increase in operational efficiencies and productivity gains in their onboarding and client management processes, mainly through the elimination of and reduction in duplicative work and manual steps[8].

Traditional Entity Resolution Process

ER is a central part of the KYC/AML process. The process provides a reliable golden record of a client that an institution is onboarding and/or maintaining a client identity during the client's lifecycle. This is important for tasks such as risk scoring through to regulatory compliance. The ER process involves the matching of entities against diverse stores of information such as internal datastores to external sets such as Companies House, or vendor-sourced data. This information may be incomplete, contain duplicates or be inconsistent resulting in a significant amount of time spent in manual review. This same review process is also observed with other data types such as name/address data and individuals. In the manual review phase corrections in terms of names and differences in addresses are accounted for before the client details can be persisted in the counter party persistence layer like a CRM or a legal entity master as summarised in Figure 4.

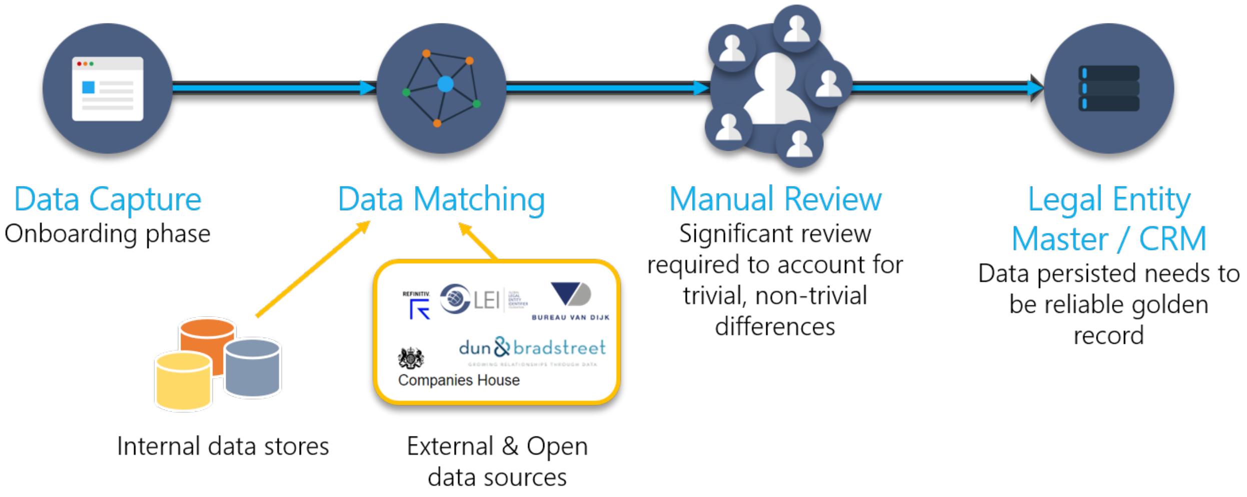

Data Capture
Onboarding phase
Data Matching
Manual Review
Significant review required to account for trivial, non-trivial differences
Legal Entity Master / CRM
Data persisted needs to be reliable golden record
REFINITIV
LEI
BUREAU VAN DIJK
dun & bradstreet
Companies House
Internal data stores
External & Open data sources

Figure 4. Summary of steps in a typical onboarding process

We view the value of using ML within the ER process to address the inefficiencies around manual tasks and improve data consistency. Below, we present research we have undertaken to measure what kind of operational efficiencies can you get from using ML in ER.

Entity Resolution Use Case Objective

Using two open entity data sources, we matched entities from the Refinitiv PermID[9] dataset against entities in the GLEIF[10] dataset. An ML classifier was applied to make predictions on the resolution of low confidence entity matches. Within this study we observed a reduction of 45% in the manual review process using ML. This frees up clients and their employees to work on more high value tasks. In addition, data consistency also improved through the ML assisted predictions.

> ***"We observed a reduction of 45% in the manual review process using ML, saving over 120 hours of manual work."***

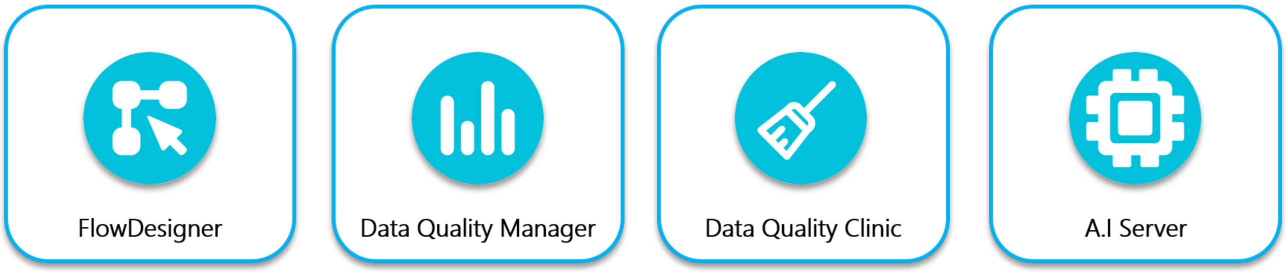
Open Entity Datasets

The two open entity data sources we used were Open PermID from Refinitiv and the LEI level 1 concatenated file from GLEIF. PermID is an open dataset from Refinitiv which provides unique ids for entities such as corporations, instruments, and individuals. The Global Legal Entity Identifier Foundation is a supra-national not-for-profit organization headquartered in Basel, Switzerland. It provides an open global resource of standardised and unique legal entity reference data. At the time of this study, the PermID dataset had 3.8M records and the GLEIF 1.4M records. We matched 100K random records from the PermID dataset to the GLEIF dataset using tools in the Datactics platform, the process is described below.

Datactics Platform Tools

We applied tools from the Datactics platform to undertake the entity matching task. Figure 5 provides an overview of the various tools and their function. Deterministic match rules were constructed using the rules platform **Flow Designer**. Data integration and scheduling was performed using the **Data**

Quality Manager. Manual review of low confidence entity matches was performed using the **Data Quality Clinic.** In the background, the **Datactics AI server** continually listened, updated, learned, and served predictions to the users.



FlowDesigner

- Drag-and-drop rules studio
- Build & test rules
- Implement business requirements
- Create, refine and test DQ processes
- Designed for business users, no programming skill required

Data Quality Manager

- Run rules
- Combine, scale and link projects
- Automate and schedule projects
- Handling of multiple data formats e.g. CSV, XML, JSON, XSLX, XLS, etc.

Data Quality Clinic

- Review of low confidence matches
- Assignment & categorisation of resolutions
- Creation of merged golden record
- Audit trail

A.I Server

- Reasoning fed from Data Quality Clinic
- Models predict matches, breaks in unseen records
- Results rendered in off-the-shelf visualisation software e.g. PowerBI
- Significantly reduces review time in manual review of breaks & failing records

Figure 5. Overview of Datactics Tooling applied in the ER Use Case process

ML Assisted Entity Resolution

In this Use Case, a Datactics Data Engineer (representing a subject matter expert in an institution) designed and developed deterministic entity match rules using the Flow Designer rules platform. These deterministic rules were used in the first phase to match the 100K PermID entities against the 1.4M entities in the GLEIF dataset as illustrated in Figure 6.

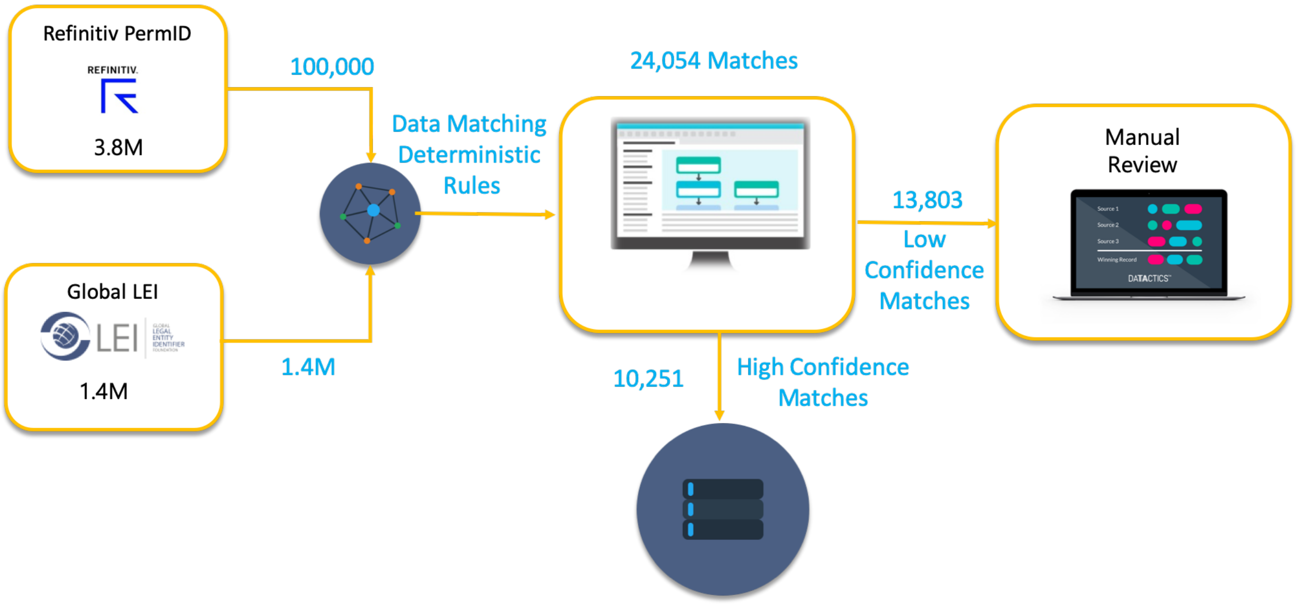


Figure 6. Illustration of the deterministic match process which is applied in the ER Use Case process

Using the rules, we obtained 24K matches between the PermID and GLEIF datasets. Around 10K of these matched were considered high confidence, meaning they scored high on the match rules. The remaining 13K match records were deemed low confidence meaning they did not reach the conservative threshold as set by the rules for a "certain" match. These low confidence matches went into manual review using the Data Quality Clinic. Manual review was performed by 6 data engineers where low confidence match pairs were reviewed to determine if there was a match, no match or uncertain match between the records. This is a time consuming and subjective process. It could take up to a 1.5mins to manually check if there was a match between two entities. To manually resolve all 13K low confidence matches would take 288 person hours to resolve. This is where we view the value of ML to aid in addressing the inefficiencies around manual processes.

Continual Machine Learning with Human-in-the-loop

A ML model was applied for the task of predicting if a low confidence match pair between a PermID and GLEIF record was a match or not. The ML model was trained using labelled examples of matches and non matches along with features engineering from the meta match data. The model learns the underlying patterns from the training data and applies this knowledge to make predictions on new cases. These predictions are presented to the reviewer performing the manual review. This is known as **continual learning with Human-in-the-loop** where the reviewer can validate the predictions and feed this information back into the model. There is a continuous feedback loop between reviewer and the model with the ML model becoming more refined and accurate over time.

Results

Figure 7A summarises a confusion matrix which shows the predictive performance of the ML model when applied to a test dataset of 2616 low confidence matches. This is data that the ML model has not seen before. We can see that this classifier was good at correctly predicting true entity matches (978) and non matches (1592). High evaluation metric results were achieved as illustrated in Figure 7B.

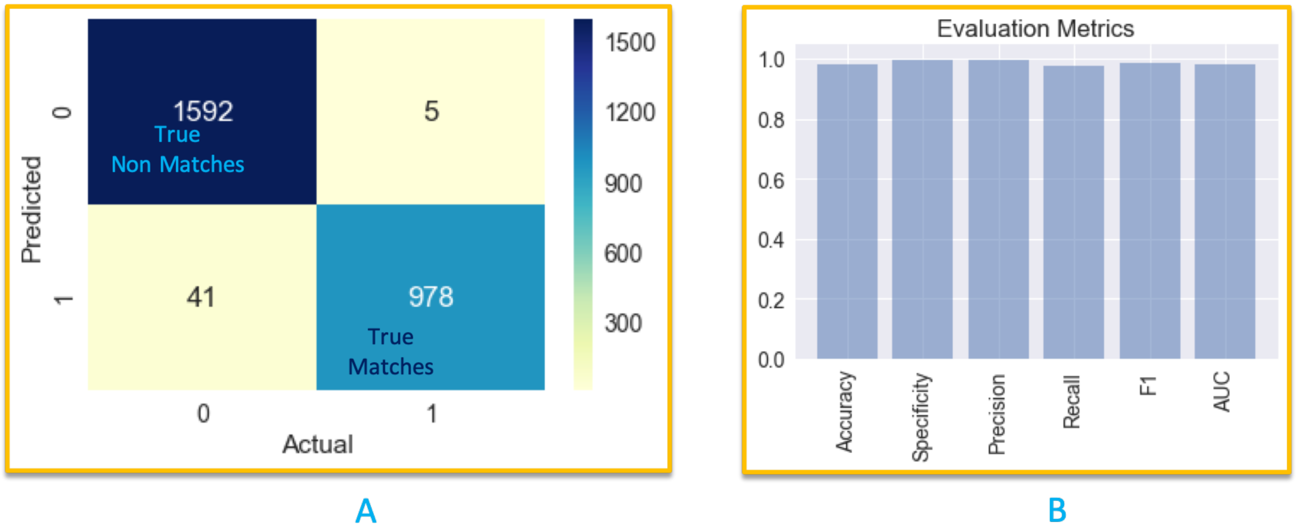
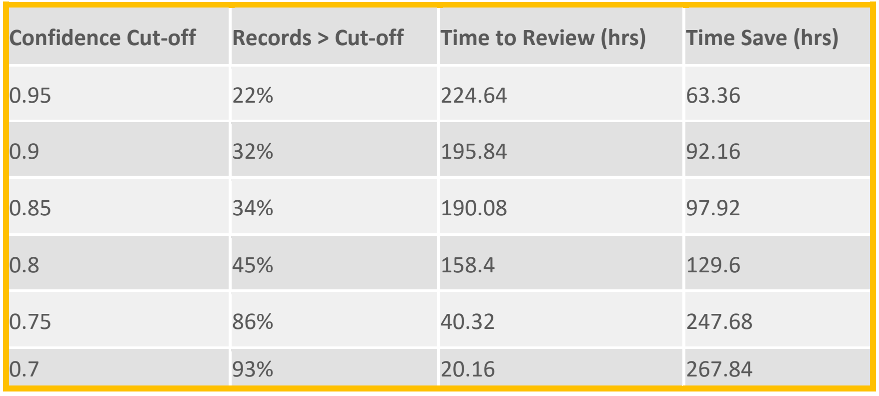


Figure 7. A. Confusion matrix presenting the classification results using a test dataset. B. Graph of the different evaluation metrics showing the classification performance.

Furthermore, we can use high confidence predictions to reduce the volume of manual review performed saving time and improving the accuracy of the process. These confidence thresholds are easily configurable, and selection will depend on the client. For example, as presented in Table 1, by selecting a confidence threshold of 0.9 and above, we can reduce the manual review by 32% and saving over 3 days of manual review.

Table 1: Summary of different confidence cut-off thresholds and the impact this has on review volume



Confidence Cut-off	Records > Cut-off	Time to Review (hrs)	Time Save (hrs)
0.95	22%	224.64	63.36
0.9	32%	195.84	92.16
0.85	34%	190.08	97.92
0.8	45%	158.4	129.6
0.75	86%	40.32	247.68
0.7	93%	20.16	267.84

This highlights how ML can be applied to work on manual time-consuming tasks enabling clients to focus on more complex match cases and enabling them to do more in their role.

5. Conclusion and Future Work

In this paper we have highlighted how the Datactics platform can be used to prepare data to be used as input to the ML Models. The platform tools can be used to perform tasks such as cleaning, deduplication, and pre-processing. Dataset features can be engineered, and labelling performed using Flow Designer and the Data Quality Clinic. Furthermore, we have highlighted how we are applying **transparent** and **explainable** ML solutions with **human in the loop** using a real-world open Use Case for Entity Resolution. By using **high confidence** predictions from ML models, we have presented manual review tasks could be **speed up** along with improvement in **data consistency** in the process.

6. References

1. Moore, S. (2018). *Smarter With Gartner*. [online] Gartner.com. Available at: https://www.gartner.com/smarterwithgartner/how-to-stop-data-quality-undermining-your-business/
2. Forrester.com. (2019a). *RQ 2.0: Assess Your Readiness For Artificial Intelligence, Automation, And Robotics*. [online] Available at: https://www.forrester.com/report/RQ+20+Assess+Your+Readiness+For+Artificial+Intelligence+Automation+And+Robotics/-/E-RES142612
3. Forrester.com. (2019b). *Your Business Is Only As Fast As Your Data*. [online] Available at: https://www.forrester.com/report/Your+Business+Is+Only+As+Fast+As+Your+Data/-/E-RES83281
4. Columbus, L. (2019). Roundup Of Machine Learning Forecasts And Market Estimates For 2019. *Forbes*. [online] 3 Apr. Available at: https://www.forbes.com/sites/louiscolumbus/2019/03/27/roundup-of-machine-learning-forecasts-and-market-estimates-2019/#4c2d416a7695
5. Deepmind. (2018). *AlphaGo Zero: Starting from scratch*. [online] Available at: https://deepmind.com/blog/article/alphago-zero-starting-scratch.
6. Nisbet, R. Miner, G. Yale, D.D.S. (2018). *Handbook of Statistical Analysis and Data Mining Applications (Second Edition).* Chapter 18 – A Data Preparation Cookbook. Available at: https://www.sciencedirect.com/science/article/pii/B9780124166325000189
7. Refinitiv. (2019). *KYC compliance — the rising challenge for financial institutions and corporates*. [online] Available at: https://www.refinitiv.com/en/resources/infographics/kyc-compliance-the-rising-challenge-for-financial-institutions
8. Forrester. (2019). *Forrester*. [online] Available at: https://go.forrester.com/
9. Refinitiv.com. (2020). *Open PermID | DEVELOPER COMMUNITY*. [online] Available at: https://developers.refinitiv.com/open-permid
10. Global LEI Foundation (2020). Dataset access information found at: https://www.gleif.org/en/lei-data/access-and-use-lei-data

Datactics Ltd.
1 Lanyon Quay
BT1 3LG
Belfast
+44 (0)28 9023 3900
datactics.com

DATACTICS™

AGILE DATA QUALITY

POWERFUL DATA MATCHING