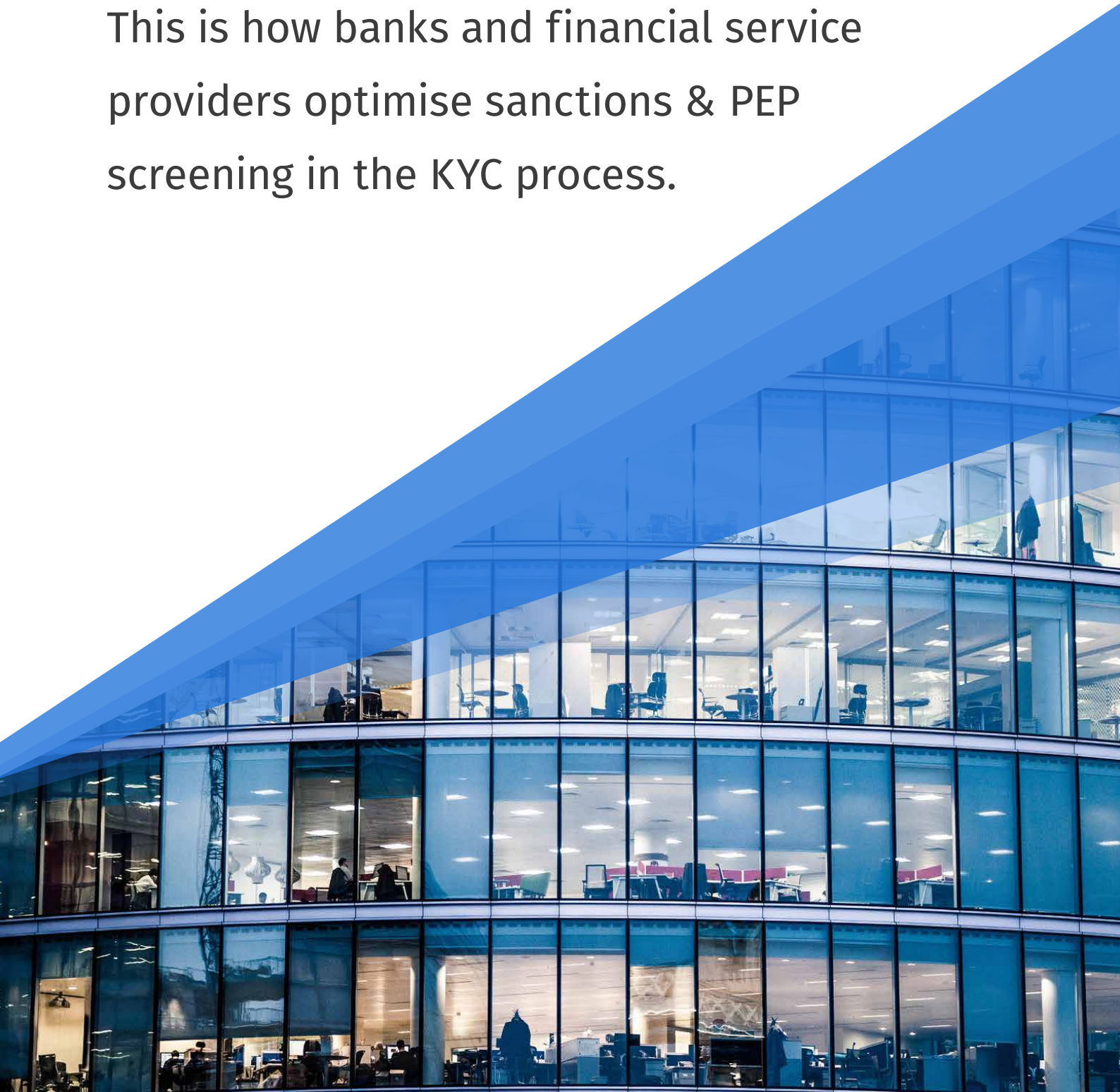
actico
Smarter Decisions

d-fine

White paper

Machine Learning in Compliance: This is how banks and financial service providers optimise sanctions & PEP screening in the KYC process.



CONTENTS

_01 Introduction

Over the past two decades, the compliance function has clearly gained in importance for banks. This is partially against a background of numerous money-laundering scandals and embargo/sanction regimes.

In an attempt to prevent money laundering and the financing of terrorism, various quantitatively oriented approaches have become established, especially the risk-based approach, which was recommended by the Financial Action Task Force (FATF) as early as 2007 and implemented in the banking sector in 2014.

TECHNICAL/QUANTITATIVE ORIENTATION DURING CUSTOMER SCREENING

The increasingly technical/quantitative orientation of the processes to prevent money laundering thus also permits the use of advanced methods and analysis options such as machine learning (ML). A simple example is the use of ML methods within the context of customer screening as required by money-laundering regulations.

CHECKING AGAINST SANCTION AND PEP LISTS IN THE KYC PROCESS

The identification and ongoing monitoring of new and existing customers as part of a so-called KYC (Know Your Customer/Client) process is a central component of these money-laundering regulations. One component of the KYC process is name matching using various lists such as sanction lists, embargo lists, PEP lists (PEP refers to politically exposed persons) and, where applicable, institution-specific blacklists.

GAINING TIME FOR TRUE MATCHES, REDUCING FALSE POSITIVES

Name matching often results in a large number of false positives. As the corresponding operators of a compliance division have to process all the positives of a matching run, these false positives lead to increased costs while minimising the time available to analyse the true positives.

HIGHER PRODUCTIVITY OF OPERATORS

By using ML it is possible, for example, clearly to reduce the number of false positives, thus reducing the work load of compliance workers and increasing their productivity. How this can be achieved in practice is shown in the following sections.

WHAT AWAITS YOU IN THE WHITE PAPER

Section 2 provides a brief overview of ML, before the actual approaches relevant for name matching application are briefly presented in Section 3. A real-life application and the results of the linked project are then discussed. The white paper ends with a brief summary.

_02 Overview of machine learning

Artificial intelligence (AI) is an established research field, with the first work on artificial neuronal networks being carried out in the 1940s. However, the "Summer Research Project on Artificial Intelligence" that took place at Dartmouth College in Hanover, USA, in 1956 is most frequently regarded as the birth of artificial intelligence.

The current great interest in the concepts and methods of AI can be explained by the following developments in recent years:

- AI applications now benefit from a large number of freely available (open source) tool kits and libraries.
- The storage capacity and computer power of modern computers and Cloud providers make the performant implementation of AI methods possible.
- The large volume and availability of data permit the efficient use of AI approaches, for example to train artificial neuronal networks.

MACHINE LEARNING AS A BRANCH OF ARTIFICIAL INTELLIGENCE

The term "machine learning" (ML) as a branch of AI describes methods that make use of learning processes to identify contexts in data sets on which predictions can be based [Murphy2012]. We can distinguish between three different ML approaches:

1. Unsupervised Learning
2. Supervised Learning
3. Reinforcement Learning

UNSUPERVISED LEARNING

Within the context of unsupervised learning, attempts are made to detect patterns in existing data sets and to derive categories from them. In this case pattern detection is not specified, but the algorithm autonomously categorises and clusters the data sets. The K-Means algorithm and the Latent Dirichlet analysis are well-known algorithms.

SUPERVISED LEARNING

Algorithms are trained on the basis of categorised data sets as part of supervised learning. The training success is reviewed with the aid of a test data set in order to evaluate the quality of the trained model/ algorithm. Actual learning takes place with the training data set, while the evaluation of the trained model is carried out with the aid of a test data set.

REINFORCEMENT LEARNING

Reinforcement learning simulates human learning behaviour. An agent independently learns a strategy in order to maximise a reward/profit. This mostly makes use of temporal difference learning algorithms known as Q-learning methods. In this method, Q refers to the benefit as a function of a state and an action.

In the application discussed in Section 4, i.e. "Name Matching Customer", supervised learning is particularly important. Specifically, observation will be used to train and implement a random forest algorithm.

_03 Random Forests and Decision Trees

Random forests can also be regarded as a set of decision trees. We will thus start by looking at the concept of decision trees.

DECISION TREES

Decision trees are used in regression and classification problems. As this article is about the "Name Matching Customer" application, this sub-section will only deal with an explanation of decision trees.

The aim of decision trees is to group or subdivide an existing data set using hierarchical decisions. The easiest decision tree consists of a node and two leaves. The node contains a logical binary rule that clearly allocates the data to which the decision tree is applied to one of the two leaves. One leaf of a decision tree can thus be seen as a response to the preceding decision. By way of an example, Figure 1 shows a group of decision trees T_i created from subsets of a data set. The results of the binary decisions are shown in colour.

STATISTICAL ALGORITHMS TO SELECT ATTRIBUTES ON THE BASIS OF THE INFORMATION CONTENT

In this application, the challenge lies in determining suitable attributes that enable classification by means of a decision rule. It is often very difficult to specify a decision rule explicitly, thus statistical algorithms are used. One of the best-known algorithms is ID3 (Iterative Dichotomiser 3) and its further development C4.5 [Quinlan1986, Quinlan1993]. The core idea of the algorithm is the selection of an attribute a based on the information content. The information gain $IG(M,a)$ of an attribute is the difference in the entropy $s(M)$ of the underlying data set M and the average entropy $s(M|a)$ for the fixed selection of the attribute a. Each additional selection of an attribute increases the size of the decision tree. There are also other statistical procedures which, however, are less suitable for the current application example, Name Matching Customer (NMC). These procedures are based on the sum of squares of the residues (Residual Sum of Squares, RSS). A binary decision to divide the data set M_i of the set M of all data at point c into the leaves $B_1(i,x)=\{M|M_i<c\}$ and $B_2(i,x)=\{M|M_i>c\}$ is optimised by minimising the sum of the residues of the two leaves for the number of all data sets i and point c. This strategy can be recursively applied to each newly created subset until a tree structure emerges [JamesWitten2017].

INCREASING ACCURACY BY REDUCING THE COMPLEXITY OF THE DECISION TREE

The depth of decision trees and thus also the degree of detail of the decisions can be limited by specifying a lower limit for allocation to a final sub-category. The optimal lower limit of a decision tree is determined by using so-called pruning procedures [BreslowAha97]. These procedures were developed to create decision trees that are not excessively adapted to the training data set used (overfitted). This means that when using pruning procedures the accuracy of correct allocation increases, as the complexity is reduced and the decision tree is simplified.

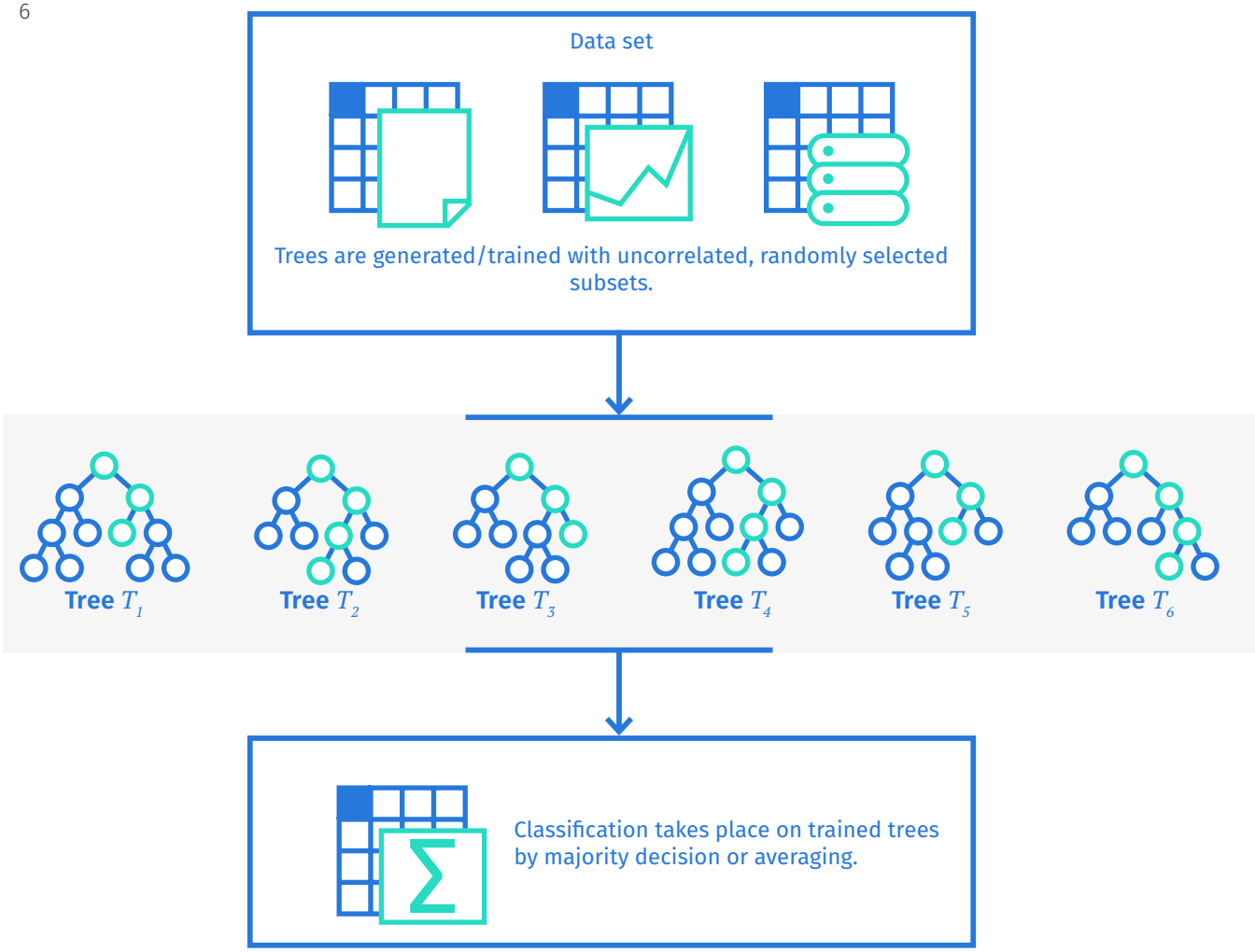


Fig. 1 Example of an ensemble of randomly generated decision trees. The aggregation logic with which a classification is made by majority decision/averaging the result of individual decision trees is shown in colour.

LINKING DECISION TREES TO HUMAN DECISIONS WITHOUT MATHEMATICAL EXPERT KNOWLEDGE

The advantages of decision trees are as follows: Decision trees are easy to visualise and to comprehend, and can be linked to human decisions without requiring mathematical expert knowledge. Decision trees are an easy way of structuring large, opaque, granular data volumes in a logical and comprehensible way, so that a granular, quantitative level is replaced with a simpler, more accessible, qualitative decision level. However, traditional decision trees often have their limits, especially when the rules generated for the nodes are very sensitive to the input data used, as this will affect the stability of the predictive accuracy. A decision depends a great deal on the distribution of the input data. If the distribution changes, the tree may become unstable, as a subsequent correction of the tree hierarchy is usually not possible without generating the entire tree anew.

One way of increasing the stability of a classification is to combine various models, in this case decision trees, or to find an average between them (Bootstrap Aggregation or Bagging). In this case each model provides a result for a random sample or subset from the total data. The individual results can also be weighted based on the size of a sample, for example. This generates rather more stable predictions for the overall classification, as averaging the individual results reduces variance.

RANDOM FORESTS

The case of a simple decision tree was explained in the previous section. Data anomalies and distributions may, however, result in a too specific categorisation by a single decision tree. This special categorisation may function for the given data base, but could fail when new data sets are added.

RANDOM FOREST MODELS COMBINE SEVERAL DECISION TREES

This restriction can be monitored by using random forest models. Random forests are based on the idea of combining several decision trees, see Figure 1. It is important to remember, though, that the decision trees should not be correlated with each other. Thus, individual decision trees are generated on the basis of randomly selected subsets of the whole data set. A randomly selected sample of the original data has the advantage that it is not the most prominent data category that is used, but that smaller categories may also be strongly represented in the random sample and can therefore be more closely integrated into the classification [JamesWitten2017]. The quality of random forests and decision trees can be described with the out-of-bag error (OOB error). This means that some of the data that have not been taken into account when creating the decision tree – out-of-bag – are used to find errors in the prediction of the correct classification.

A standard method to describe the quality of a decision tree is the use of an entropy function,

$$s = \sum_{j=1}^{M} p_j log(p_j).$$

where p_j is the probability with which a data set M_i has been allocated to a classification j. Entropy is minimal when all data in one class coincide. The entropy of a binary tree, consisting of a root and two leaves, is maximised when the data to be classified are equally allocated to the two leaves.

An alternative approach to determine the quality or impurity of a decision tree is the Gini coefficient. In practice, the Gini coefficient and entropy usually produce very similar results, so that it is generally sufficient to use only a single impurity criterion.

RANDOM FORESTS ARE PARTICULARLY WELL SUITED TO COMPLIANCE

In the compliance environment, the underlying data may rapidly change. Thus, customer data and checklists, for example those containing criminal, well-known or politically exposed persons, may yield very good results in an existing model. However, when updating the data base, there is a risk that an original decision tree will no longer lead to the desired result of a meaningful classification. This means that random forests are particularly suitable for use in the compliance environment. They constitute an important tool for creating statistical correlations between data sets, thereby supporting the monitoring of new and existing customers with standardised processes. Randomly created decision hierarchies always have a certain inaccuracy and may also make use of logically false correlations.

HOW COMPLIANCE OPERATORS IMPROVE THE CLASSIFICATION

ACTICO provides a closed Compliance Suite for this purpose, which can be fully integrated into existing systems. This always provides the operator with the option of checking an individual classification and using manual feedback iteratively to improve future classification.

The next section will deal with the functional details and the functionality of the ACTICO software *Name Matching Customer*, based on the background that has been theoretically discussed before.

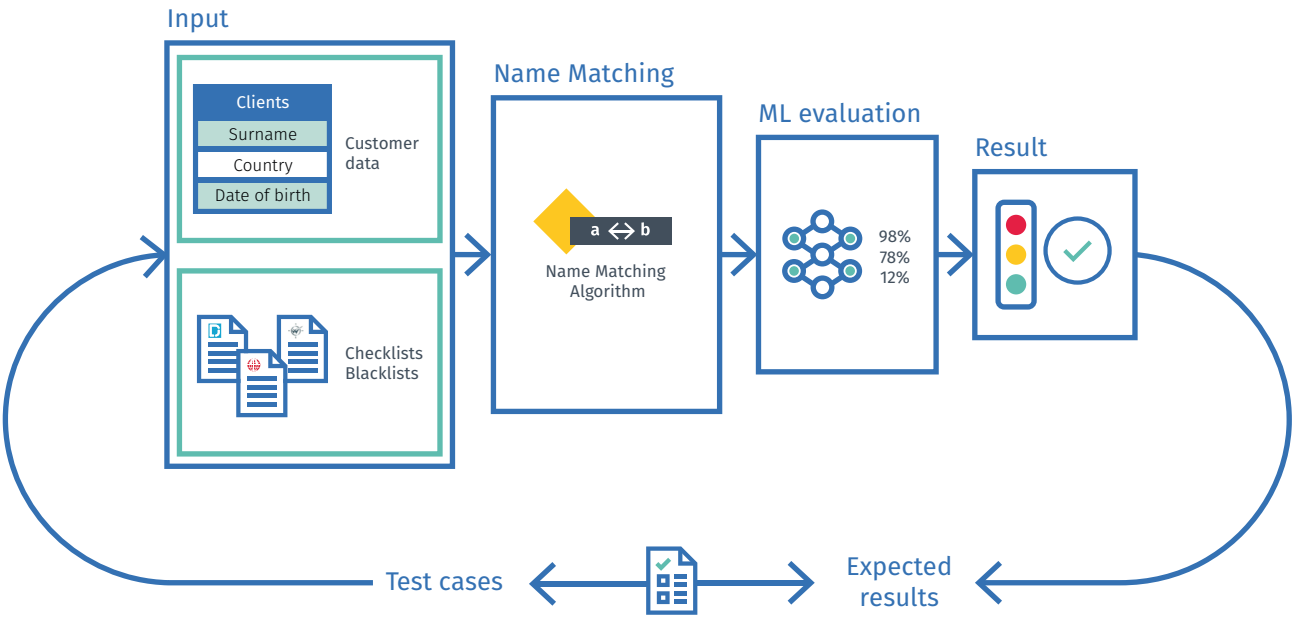
_04 Application example for the comparison of customer data to checklists with the aid of machine learning

This chapter shows how machine learning (ML) can be used in compliance. The example shown is name matching, using the software module *Name Matching Customer (NMC)* of the ACTICO Compliance Suite. Machine learning makes use of the tools of the Compliance Suite machine learning module. Among other things it supports the strategy of supervised learning with random forests described above, as well as other algorithms such as deep learning, which makes use of neuronal networks.

SAFE KYC PROCESS BY SCREENING CUSTOMER DATA

The *Name Matching Customer* module of the Compliance Suite compares customer data against checklists. A person may be listed for various reasons. The lists may contain the names of criminals and terrorists, as well as people with political influence (politically exposed persons, PEPs). For potential new customers, comparison takes place as part of the KYC or Client Due Diligence (CDD) process before starting business relations.

For existing customers, comparison takes place on a regular basis following relevant changes to the customer master data or listed persons. The comparative algorithm makes use of names, countries (domiciles, nationalities) and dates of birth to find possible similarities. For names there is also a fuzzy matching (similar names). If a potential match is found, this is clarified in the software, as an operator documents whether this is a true match.

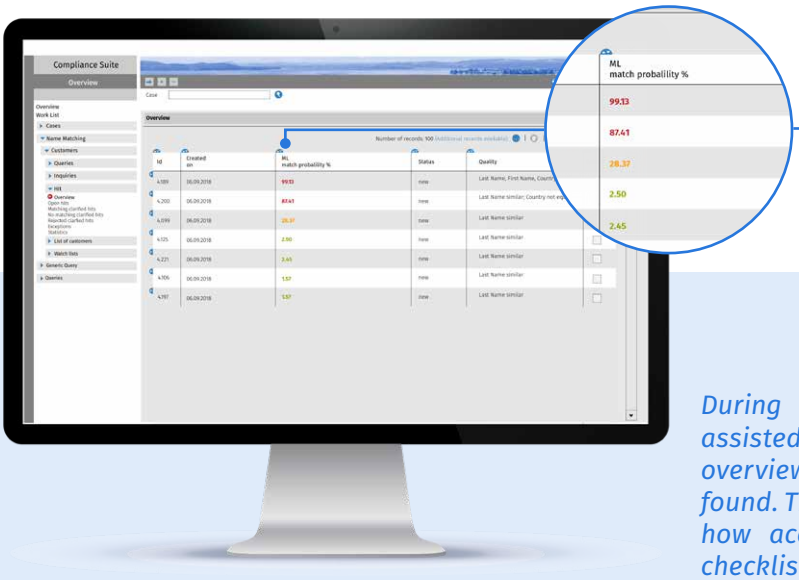


The software checks new and existing customers for matches with entries in sanction and PEP lists. Machine learning evaluates the matches with the aid of a learned model.

Fig. 2 Name matching supplemented by machine learning evaluation.

FOCUSING ON TRUE POSITIVES

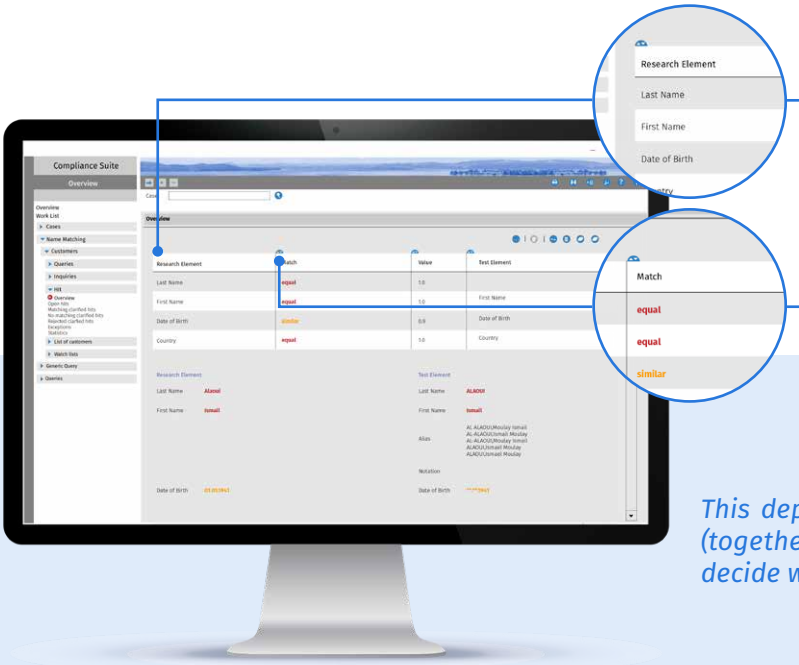
This software has been used for years by many customers. In the meantime, the comparative algorithm has been optimised to the degree that, where possible, all real matches (true positives) are found, but that nevertheless the lowest possible number of non-corresponding matches (false positives) are generated. This optimisation is currently being further improved by machine learning. Comparison using the algorithm is followed by automatic evaluation with a learned model. This predicts the probability of a possible match being documented as a true positive during clarification as well, thus allowing the prioritised clarification of the possible matches.



Match probability in %
Quality of match features between customer data and sanction list entries.

During clarification, the employee is assisted by being presented with an overview of the full and fuzzy matches found. This shows which data correspond how accurately with an entry in the checklist.

Fig. 3 Listing the check results for the operator, prioritised according to their ML evaluation.



This process shows the investigated elements such as the surname, first name and date of birth.

The operator can see which elements are identical or similar to entries on the sanction list.

This depiction is used by the processor (together with additional research) to decide whether there is a true match.

Fig. 4 Presentation of a potential match for clarification by the operator.

LEARNING A MODEL FOR NAME MATCHING

In the case of Name Matching Customer, users of the software have already clarified a number of potential matches. The result of the clarification has been documented in the system database along with other data relevant for the case.

The following diagram shows how a model can be learned and implemented from the characteristics of existing cases:

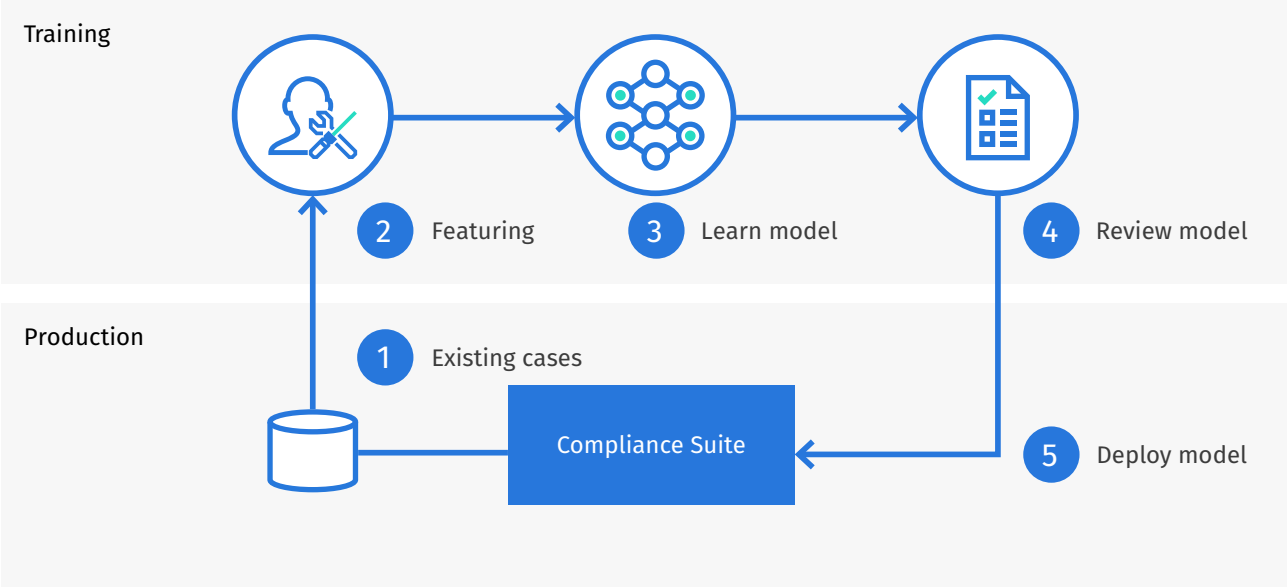


Fig. 5 Generating a model with the aid of machine learning in the ACTICO Compliance Suite.

1. Feature engineering is used to extract the features and the expected result (label) from the cases with a result contained in the database.
2. The extracted data are divided into training data and test data.
3. Supervised learning methods are then used to generate models from the training data.
4. The test data are used to test the models.
5. Following a review, models can be taken into operation.

The features used for the ML model are as follows:

- The information about which comparisons have been made by the algorithm, for example a comparison of the customer's surname with the surnames on the list or a comparison of the first name and surname of the customer with an alias on the list.
- The information as to which result was used to carry out the comparison, for example a full match, fuzzy match, etc.

The actual customer data and list entries are not part of the features. This is done for data protection purposes, so that the features do not contain any information that could be used to identify the customer.

EVALUATION OF THE LEARNED MODEL

The evaluation of a model may take place in various ways. When making a division into two classes, as shown here, this can be displayed as a limit value optimisation curve (receiver operating characteristic, ROC).

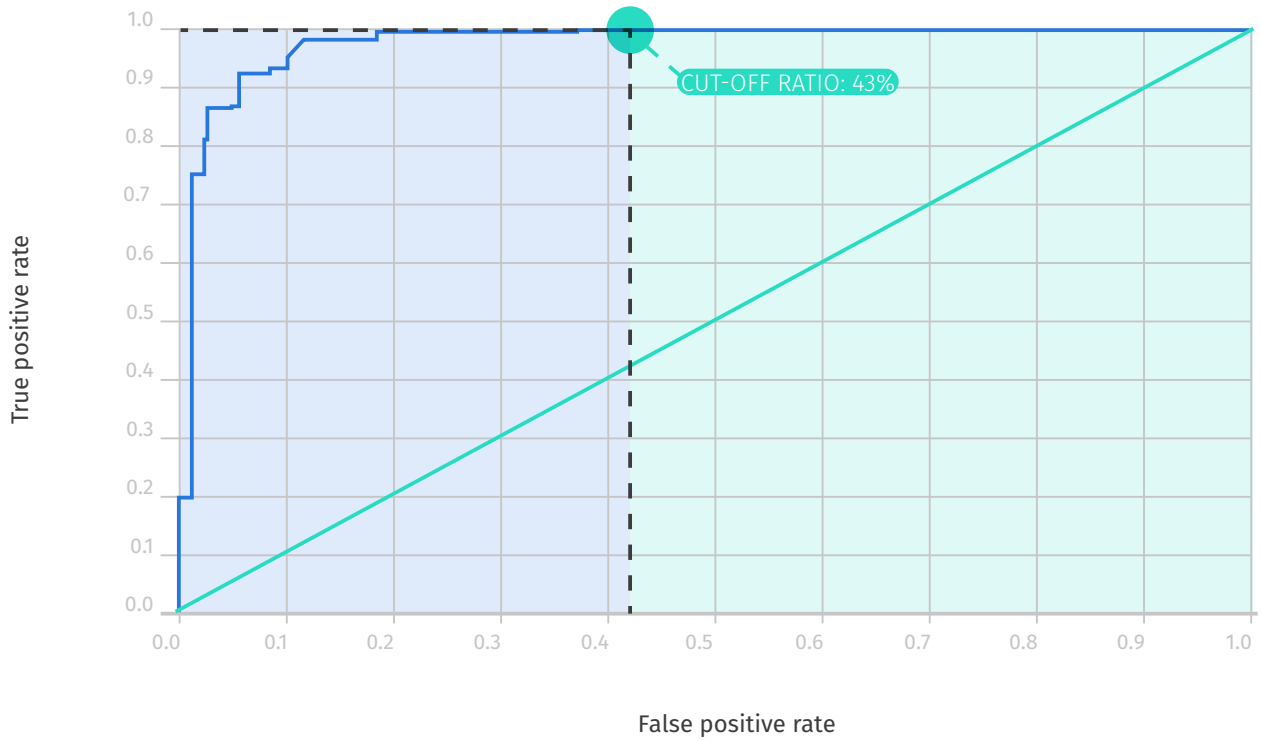


Fig. 6 By using machine learning, compliance divisions save approximately 57% of their clarifications.

Saving up to **57 %** of the clarifications in the KYC process.

The more the blue curve in the diagram deviates towards the top left from the diagonal, the better the cases can be classified. If, for this example, cases are prioritised according to the probability of a match as determined by ML, no further true positive is found after approximately 43% of the cases.

In practice, models were first generated using the data of six customers. Over 25,000 data sets were available in all cases. Random forest was thus used to generate models that classify the cases well. Normally, 30% to 40% of the clarifications can be reliably categorised as non-matches. In individual cases (as above), this figure may even be higher.

_05 Summary

The use of ML techniques in the financial industry, especially in risk controlling and in the compliance division of banks, now comprises a wide range of applications. In this article we have focused on so-called supervised learning with random forests and highlighted a case study taken from the compliance environment and based on this method in more detail. In this case the use of ML can result in a classification or prioritisation of name matching hits. By identifying false positives and excluding these non-relevant matches from subsequent processing the costs of match analysis can clearly be reduced.

The customer screening process discussed in this article is only one potential application of AI or ML approaches in the compliance sector. There have also been attempts to monitor the reputation risk by handling compliance risks using an analysis of unstructured communication data [DobrikovGraf2019].

The analysis of communication data using ML methods can also be applied to far more complex scenarios, such as fraud prevention and the avoidance of insider trading. In addition to compliance-related fields, there are also applications dealing with credit risk monitoring [DobrikovGraf2017].

_06 Sources

[Bishop2006] C. Bishop, Pattern Recognition and Machine Learning. Springer, 2006

[BreslowAha97] L. A. Breslow and D. W. Aha, Simplifying Decision Trees: A Survey, The Knowledge Engineering Review, Vol 12 (1), 1997.

[Buxmann2019] P. Buxmann, H.Schmidt (eds.), Künstliche Intelligenz. Springer Gabler, 2019

[Murphy2012] K.P. Murphy, Machine learning: a probabilistic perspective. MIT Press, 2012

[Raschka2018] S. Raschka, V. Mirjalili, Machine Learning mit Python und Scikit-Learn und TensorFlow. mitp, 2018

[DobrikovGraf2017] T. Dobrikov und F. Graf, Nachrichten in Frühwarnsystemen und dem Kreditrisikomanagement, Zeitschrift für das gesamte Kreditwesen, Heft 09/2017

[DobrikovGraf2019] T. Dobrikov, F. Graf, S. Stadelmann, S. Ulsamer, Kontrolle des Reputationsrisikos: Management von Compliance-Risiken durch Analyse unstrukturierter Kommunikationsdaten [Management of compliance risks by analysing unstructured communication data], FIRM Annual 2019

[JamesWitten2017] G. James, D. Witten, T. Hasti, and R. Tibshirani, An Introduction to Statistical Learning, Springer New York Heidelberg Dordrecht London, 2017 DOI 10.1007/978-1-4614-7138-7

[Quinlan1986] Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81–106

[Quinlan1993] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

AUTHORS
d-fine:
Dr Ulrich Lechner
Dr Marcel Langenberg

ACTICO:
Thomas Ohlemacher

_07 List of figures

d-fine is a European consulting firm focused on analytical and quantitative challenges and the development of sustainable technological solutions. The combination of hundreds of employees with a strong scientific background and many years of practical experience enables us to provide tailor-made, efficient and reliable solutions for more than two hundred clients from every sector of the economy.

More information at: www.d-fine.com

ACTICO is a leading international provider of software for intelligent automation and digital decisioning. Its scalable software combines rule-based technology and machine learning in a unique way and is totally audit-proof. It allows companies of all sizes to process huge amounts of data and make and automate AI-supported and rule-based decisions in real time. Through intelligent automation, ACTICO increases the business value of its customers by improving their operational decisions.

The company was originally set up as Innovations Software Technology, but it has been trading as ACTICO since 2015. Today, its customers include leading companies in 35+ countries, including Volkswagen Financial Services, ING and KfW. The company employs 140 people worldwide and has offices in Chicago and Singapore. Its head office is located in Immenstaad, Germany.

More information at www.actico.com

ACTICO

Europe
ACTICO GmbH
Ziegelei 5
88090 Immenstaad
Germany

Americas
ACTICO Corp.
141 W. Jackson Blvd.
Ste 300A
Chicago, IL 60604
USA

Asia & Pacific
ACTICO Pte. Ltd.
#11 - 04, The Arcade
11 Collyer Quay
049317 Singapore

info@actico.com
www.actico.com

actico
Smarter Decisions

04/2020