Deep Science Publishing
https://doi.org/10.70593/978-93-49910-38-6_7

DeepScience
Open Access Books

Chapter 7: Risk and compliance analytics using machine learning

7.1. Introduction to Risk and Compliance Analytics

Risk is an essential concept in all walks of life. All decisions contain an inherent level of risk. One can choose to be exposed to and exploit risk or choose to avoid risk and thus (presumably) yield lower expected returns. However, both the risk-averse and risk-seeking need clear risk analytics in order to quantitatively know where they stand. In the absence of such knowledge, even those who exploit risk might be left exposed to far greater risk than they anticipated. Similarly, the risk-averse should understand where their exposure starts and stops. The subject of risk analytics spans a wide horizon, from the estimations of the risk of exotic derivatives, such as path-dependent options in the finance arena, to the forecasting the propagation of risk in a complex network of manufacturing machines. All these pursuits are ultimately based on data; whether this is market data, sensor data or even expert knowledge, the knowledge is distilled and put into play by way of data. Data-driven modelling and analytics is a digital representation of the real world used to quantify, manage, and analyse risks. Risk projections, such as the computing and risk-piece of the picture, are then done in response to the chosen risk model. Like any analytics or data-driven effort, risk analytics also faces data challenges at each stage of the analytics workout, here outlined as data difficulties and addressed with unit operations and performance measures. Considerable frameworks for the auditing of data during risk analytics are outlined and network metrics are derived for this audit process. These solutions provided are meant as high-level guidelines, opening a conduit for academia to provide even greater granularity.

All classification models were built upon ensemble learning via original credit scoring architecture based on xgboost. The process of hyperparameter tuning was manually conducted and facilitated in order to attain a maximum accuracy level. Following global testing conducted via extended cross-validation and stratified KFold partitioning of the dataset, all models used in feedback - bucketing processing were model-agnostic, hence

novel methods were integrated into every solution stream. Such capabilities included several dimensionality reduction approaches, optimised restriction of analysed dataset size and ranges as well as two sampling methods that help tackle the class imbalance issue of banking analysis datasets. Finally, a unified module was also built for cumulative interest benefits, simulations and visualisation of the models performance metrics.

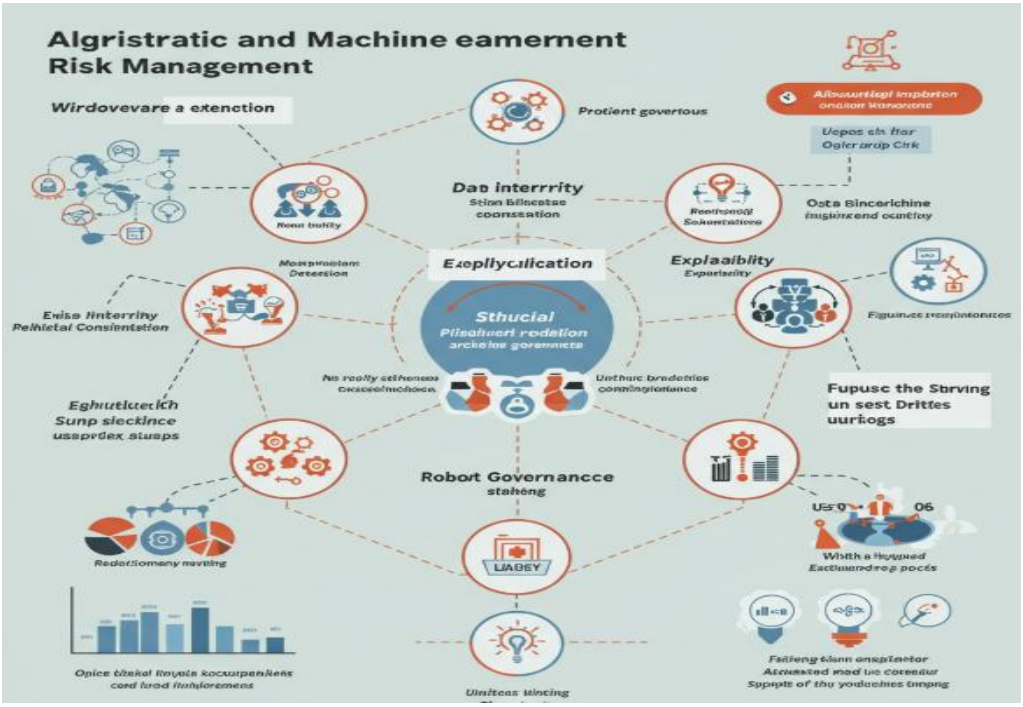

Fig 7.1: Algorithmic and machine learning risk management

7.1.1. Research design

In order to fulfil the objectives and answer the aims of the dissertation work, the design of a methodology that brings together the technical requirements and the required datasets was developed. The chosen design is a solution portfolio consisting of various solutions applicable across the dimensions of the chosen analytics field. This helps tailor and broaden the applicability of any given individual solution based on specific business needs and advanced analytics skill sets on an individual basis. The solution portfolio feeds each of the developed solutions with key pre-processing, feature and set conditioning capabilities needed to ensure meaningful results. Nevertheless, it is essential to note that while generalisation is attempted, all datasets should be treated as

individual cases and appropriate validation on the final deployment environment should be conducted on any dataset prior to using it in a production environment.

The initial analysis consists of the exploratory data analysis, a key step that should take place for all solutions. Through visualisation and insight extraction capabilities, this step will allow for the construction of a foundation necessary to better comprehend the given dataset. During the preparatory stage, several pre-processing capabilities were built to handle missing data, remove or limit irrelevant features and complements the data set with a more informative target variable. Text data handling capabilities were also built, employing various tokenization and n-gram approaches. Various pre-processing methods were tested prior to developing the solution itself, ranging from TFIDF, sentiment analysis, word embeddings and open source LLM libraries. Ultimately, a custom embeddings based document payload extraction pipeline in a vector database was developed and deployed.

The technical deliverables were developed for a low- or no-code solution deployment alongside more tailor-designed off-the-shelf executive roadmaps and presentations, as well as more generic use-case specific documentation for any potential future users. Considerations regarding the overall architecture development, deployment of the technical components with focus on ingestion and compliance side and the included technical and business visualisation dashboards, as well as overall protection, are to be carried out.

7.2. The Role of Machine Learning in Risk Management

Risk and compliance analytics involves the identification, assessment, and prioritization of risk, as well as the application of resources to minimize the impact of the uncertainty on organizational objectives (Gounagias et al., 2018; Savić et al., 2021; Aladebumoye, 2025). The priorities and resources allocated to it vary from organization to organization according to their strategies and risk appetite. There is a tension between organizations managing risk on their own, or addressing it as a collaborative operation. There is also a tension between a non-intrusive, automated approach to risk management and monitoring compliance, or an intrusive, manual and PMO-style approach to management.

Nevertheless, there have been developments in thorough and sophisticated frameworks and systems for holistic risk and compliance analytics, as well as monitoring. Machine learning (ML) has been a topic of interest among selected organizations, independent research institutions, and academia, and with the technical capabilities made available by cloud computing, ML and big data has also found its way into risk and compliance analytics systems and databases. They offer organizations enormous opportunities. The

development of these technologies shifted the technical focus of risk and compliance analytics. The market followed singular solutions focusing on isolated risk types and on just one type of data, transactional data for instance, on their own advantage without adherence to the value still residing in the interrelations between the technical channels, as well as the data silos.

The beginning of consequence-proof, holistic risk and compliance analytics platforms slightly predates the ML and big data frenzy, but in terms of advanced design and implementation it is still at its very premature stage. Risk and compliance analytics vendors focus on the visualization of and on the KYP rationale prioritization and monitoring of risks and compliance regulations. As a result, the rescheduling, removal, and revision of controls and compliance regulations as a consequence-proof design element cannot be modeled, formalized, or visualized, and consequently do not exist in the scientific and vendor literature. By consequence-proof risk and compliance analytics is the holistic capability of a platform to assess the consequences of risks or compliance object violations and by either the intended or the unintended enactment of controls.

7.2.1. Research design

As organizations increasingly rely on technology and accumulate larger volumes of data, the analysis of this data typically generates useful metrics and indicators to facilitate key decision-making processes. The term “risk analytics” encompasses several streams of business analytics, including performance analytics, marketing analytics, operational analytics, financial analytics, risk and compliance analytics, and environmental analytics. As financial institutions are particularly vulnerable to risks, they monitor varying risks and use IT systems to process large data volumes. However, regardless of the existing regulatory frameworks, governments need to ensure that regulations are sufficiently granular to mitigate systemic risk and impose capital and liquidity requirements on large institutions. Regulatory and risk management frameworks for global banks, investment banks, and insurance companies need to capture individual and collective risks stemming from an institution’s activities and portfolio.

Automatic credit scoring provides a reliable quantitative indicator for lenders’ credit data analysis and risk management. Arbitrary high default costs can cause significant losses for borderline applicants inadequately categorized. Conversely, high cash flow and overdraft limits of otherwise solvent low-risk applicants entail unnecessary lost opportunities for savings. Credit scoring enhances long-term survival in consumer credit markets and is relevant for all industries where trust is required between the provider and consumer. Machine learning and big data analysis approaches can be effectively applied to large credit data. Special attention should be paid to ensuring the trained scoring model’s interpretability, especially in the area of bank credit risk analysis.

7.3. Types of Risks Addressed by Analytics

While risk analytics is one of the most important and prominent applications of data analytics, there are other critical applications in the areas of compliance analytics, model validation analytics, market surveillance, insider trading and fraud detection, anti-money laundering, and regulatory compliance (Zheng et al., 2018; Zhang et al., 2022). Banks need to develop risk models for the lending and trading portfolios to calculate VaR and economic capital. Risk and compliance analytics initiatives in a bank cater to MDA, risk governance, compliance with regulatory mandates, and the assessment of enterprise-wide risks. Risk models such as VaR, counterparty credit risk, and stress testing run with high granularity, frequency, and multiple market scenarios.

Fig 7.2: Risk Management Types & Assessment

In addition to regulatory compliance and risk management analytics, banks conduct a range of comprehensive analytics initiatives to measure and monitor operational risks and expected and unexpected losses due to frauds and processing errors. For example, credit scoring model development and validation require mapping high-dimensional and sparse rid data into risk classes. Earning at Risk (EaR) models are developed to quantify the market risks in trading portfolios. Banks also persistently scan the news and social

media for indications of possible future credit defaults and large community-level error or frauds and conduct agents-based simulations of bank runs and liquidity crisis. AML analytics for laundering red-flag transactions usually involve combining transaction-level information with customer demographics. Banks also maintain vast historical databases on issuers, trades, and fraudulent red-flag transactions, and conduct pattern recognition analytics to develop fraud detection rules.

7.3.1. Operational Risks

Operational risk differs from other types of risk precisely by remaining unnoticed until the day that gives rise to a consequence like that of a large loss in the trading desk or a scandal affecting the bank's management. Unlike market risk, liquidity risk, loan and credit risks and all the other risks that have taken the central stage after the banking crisis, marketing risk can arise from a poorly enforced process in the organization that is not well known while situated miles away from the risk report center. The aim is to evaluate a global level of operational risk from a combination of models on the many individual sources. This turns out to be feasible, the global behavior of the system being characterized by self-organized criticality.

In any complex system losses come from two different origins: on one side losses are generated spontaneously by the individual processes, on the other side losses on one process can be triggered by a loss occurring on another process. When a trigger happens, a cascade of losses can spread like an avalanche. Also banks can be regarded as large complex systems subject to operational risk. Indeed various kinds of operational losses can occur in banks and a database composed of a number of events and damages is typically collected in a bank. In view of the high costs of this risk, banks must assess the level of operational risk to allocate a sufficient capital requirement.

Mechanisms similar to those occurring in banks were successfully investigated in the context of different systems, such as: earthquakes, solar flares or forest fires. For banks the main objectives are to try to understand if and how losses can be forecast, and to explain the power-law tail behavior characterizing the distribution of large losses. By modeling the global losses in banks as resulting from the interaction between the many sources, this turns out to be feasible and most of the exponents characterizing the distributions can be explained from the properties of individual sources.

7.3.2. Financial Risks

Financial risks refer to the current condition of a target company's finances that may result in bankruptcy, fraud, customer churn, etc. to depict various types of financial risks.

Also, assessing one's own financial condition is very important for enterprises, as those who possess high risks may incur huge losses. Financial risk prediction aims to retain the valuable potential customers while cutting off the non-potential customers shortly. On the one hand, the right customers must be retained so as to collect future cash flows and other information while increasing the market share.

7.4. Data Sources for Risk Analytics

The development of robust risk analytics capability requires processing large volumes of data from diverse internal and external sources, which can be difficult to centralise and store in a single data warehouse. Different data sources are common in the financial market. Here data source categories and examples for risk and compliance analytics are illustrated.

The analytics team obtains trading position data as well as reference data via an application programming interface (API) that connects a UI-based interface and a working database in the risk analytics engine. The front-end system deals directly with the consortium database. In this case, a local working database was built and synchronised at the end of each day with the consortium database held on the local server.

Reference data can also be updated on a daily basis in a similar manner. Depending on the data type, another application query can also be used. The graphic display of data types and sources is also illustrated in this section. Financial price data is obtained from an external data provider. Financial price data is also included in a consolidated database that contains minute intraday trading data, daily movement data, forward settlement data and correlated variable data from different sources.

Fundamental and analyst research source data is obtained using an application that spots key events, thresholds and ratios of company fundamentals. By extracting the data manually at the end of each day, key events and other processing results can be imported into the database for use in risk assessment with analysis rules. Tables with more than two million data points can be generated in less than 15 minutes. With the same treatment, analyst reports can also be processed.

The major source of alternative unstructured data is a local database where content collected regularly from different sources has been parsed. The parsed data are now granularly stored in document-level tables, which include attributes needed for meta analysis. Another scraped document table is used for collecting news articles from general media sources. Because news articles are provided in XML format, document-level attributes are directly parsed and stored in the same table.

Due to the fast-growing data volume, an automated process was implemented in which newly saved documents can be analysed periodically to directly extract keywords and build the document-keyword and keyword-graph tables. Extracted keywords for a daily topic modelling result can be visualised using a digital graph toolbox. In addition to processing, various models and analyses used to extract wellness scores from so-called big text documents have also been implemented.

7.4.1. Internal Data Sources

Examples of internal data sources include event log data produced by software applications, legacy systems or databases where "shadow" data is stored, formal reports used to store outcomes of compliance checks, or any other data stored internally which is not part of the GRC tool but could still be potentially relevant to continuous compliance monitoring (CCM) or forensics. The concept of shadow data is often specifically invoked in the context of this type of internal data source. Shadow data refers to data created by expert employees of an organization internally to address regulatory needs which is not stored formally in official channels and therefore has the potential to "go dark. The acquisition of this data type typically requires the tricky task of getting easy and conventional access to them, which is often motivated by the bright perspectives of cost-effective and reasonably accurate insights instead of comprehensive monitoring of a certain GRC area, and the selection of appropriate methodologies to automatically extract the information they contain.

An example problem in this field is regarding public procurement regulations in the European Union. The acquisition of relevant event logs was difficult since existing e-tendering systems were not easily open for software engineers coming from the Regulation-Based Compliance (RBC) community. An obligation for public contracting was opened to public bids, and event logs containing information on all invited bidders and the responsible contracting authority would be necessary for later monitoring the compliance with this rule. When engaging in automatic model extraction or event log alignment, proprietary systems hindered inspection on the event logs to develop an affordable solution for the problem. Instead, during the time leading to the data acquisition, e-tendering systems offered publicly available and legally required robust compliance reports that could assess public procurements at a high-level compliance indicator resolution.

7.4.2. External Data Sources

Relying heavily on external data sources for machine learning in official statistics comes with significant risks. Such a dependence can leave statistical agencies vulnerable as

they have limited control over these sources. A change in data types or schemas refers to modifications made to the data formats or the structure in which the data are stored and offered. These types of changes may arise due to a need to accommodate future use cases or business requirements, to eliminate technical debt, or to improve data storage and retrieval efficiency. It is important to be vigilant about changes in data types or schemas, as even seemingly minor adjustments can have significant impacts further down the data pipeline. To mitigate these risks, it is advisable to stay informed about data change announcements from providers and to implement robust data checks during data ingestion. Additionally, the deployment of effective monitoring systems can help catch machine learning failures quickly and prevent potentially costly errors. One significant approach for machine learning use cases is the reliance on external data to complement or augment the current data supply. Such a desire has manifested itself significantly with the rise in popularity of big data: massive volumes of information that often contain rich, valuable patterns and knowledge about the world. The rise in data availability is unquestionable, with external providers being prominent examples. However, an increasing dependence on these external data sources could come with risks. Such a data source can be seen as a "black box" and as there is often not much machine learning practitioners can do, agencies have to rely on their external provider to keep it up and running. As such, the following types of risks may occur: changes to data types or schemas, data availability or technical issues. Machine learning use cases can be built on fragile foundations, as one slight misstep elsewhere in the pipeline can completely ruin the quality of the results. It is crucial to respect the ML abundance and make sure that the agency is as robust as possible against the aforementioned risks.

7.5. Machine Learning Techniques for Risk Assessment

Risk classification needs to consider the trade off between risk premium and reputation cost. The higher risk a contract is, the more premium could be considered as reflected in order to ensure trustworthiness. On the contrary, untrustworthy contracts should be avoided due to potential reputation loss. Local MinMax normalization technology is applied to transform the original attributes into fractional equivalent attributes within the range of. Since ML methods respond to distance metrics, original attributes must be numerically processed. Owing to different attributes being measured in different scales, direct application of distance-based ML mechanisms could result in insensitivity of the prediction accuracy due to the existence of dominant attributes.

Lisa will be required to pay for her customers' incentive programs and phone losses. These obligations are expected to be a secret before an agreement is reached, as they are taken into account in the equivalence of bargains. For this reason, the product can be treated as a high risk contract with a low chance of success. To insult these contracts,

Lisa could either avoid eligibility for the contracts altogether or, if eligible, magnify proof of identity, credit rating, and residence in order to reject malicious behavior possibilities. This bias could be done by requiring hard verification materials for registration. Due to the privacy concerns of customers, this needs to be traded off against reputation loss and operational cost. Note that non-cooperative game theory is applicable for this case as well and grants the possibility of simultaneous optimal parameter choices for all parties. Capacities adopted in this work reflect a state-independent price mechanism with no bids placed by agents. However, it would be quite unrealistic because any one agent could communicate with others in order to produce joint value draws and significantly increase defect possibility. Bell work with higher moral standards could be studied considering differences in agent preferences.

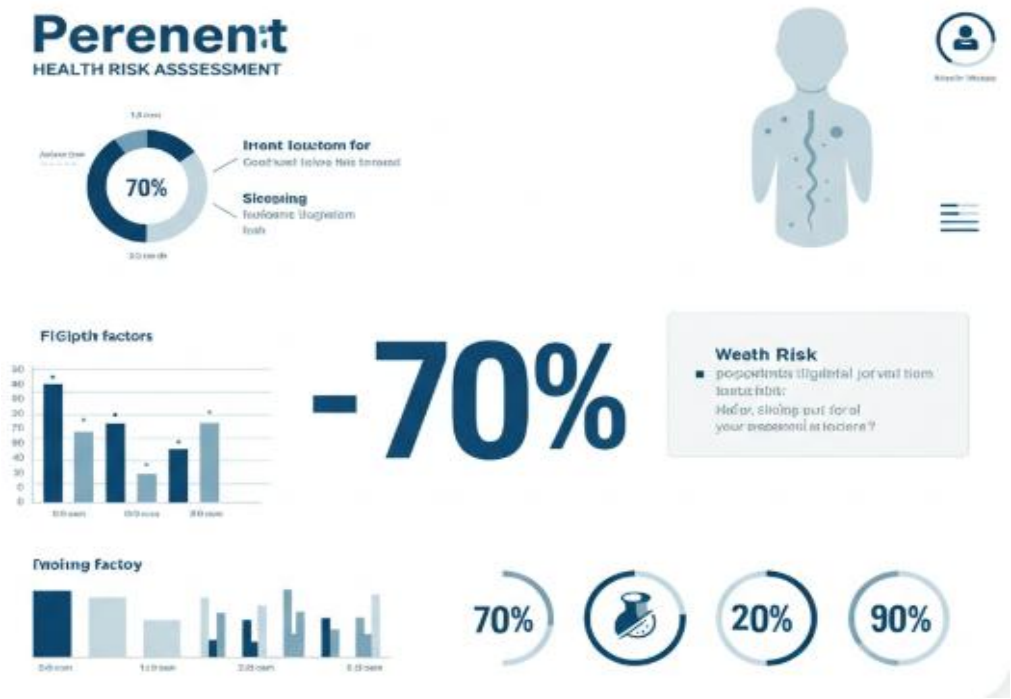

Fig : Health Risk Assessment Using Machine Learning

7.5.1. Supervised Learning

Supervised learning is a category of machine learning algorithms that require the training data to include the inputs and their corresponding expected outputs (target labels) to develop a prediction model. When the objective is to classify observations into categories, the supervised learning is known as classification. If the goal is to predict a continuous numerical target label, then it is called regression. Supervised machine

learning requires a training dataset with inputs and their associated expected outputs (prediction target). The training dataset is then partitioned into a training sample (used to train the Supervised Machine Learning models) and a testing sample (where the performance of the Supervised Machine Learning models is evaluated).

When there is only a single target variable with two possible values, the supervised learning is referred to as binary classification. For risk and compliance analytics, the EWS model supported by supervised learning for binary classification is one of the most important application areas. While some of the algorithms were presented as variants of the same HRDDM, they take advantage of the algorithms that can learn decision boundaries in different forms. They allow feature selection capability that considers univariate and multivariate relations between the inputs and the target variable. Also, some of them can perform the task in the streaming environment.

The characteristics of the applications were first introduced prior to the presentation of the algorithms. Although they differ in domain and size, the supervised learning for binary classification is generic and systematic. Based on this structured view of the applications, the processes are simplified and can be developed robustly and adaptively. Initially addressing a single application, all the components involved in the ESW models are presented and fully elaborated, together with performance analysis. Afterward, the focus was switched to greedily developed HRDDM based on naive Bayes and then filtering and boosting (Ensemble). Based on a principled systematic procedure, the proposed approaches enable the RSF to collect a rich and diverse set of candidates and reduce the risk of missing suitable algorithms.

7.5.2. Unsupervised Learning

The most commonly used type of machine learning is supervised learning, where there are training data sets with sample input-output pairs, iteratively used to adjust model parameters with the goal of better classification or regression performance, preferably assessed on separate test data. A basic supervised model created for risk and compliance analytics tasks would be a classifier. Unsupervised learning refers to a family of machine learning methods that are applied to construct models from data where output is unknown. It is often used to explore the general structure of data and identify groups, patterns, or outliers. Unsupervised methods were also applied to risk and compliance analytics tasks, but mostly in the early stages, when modeling was done by simple methods like clustering with k-means and/or calculating traditional metrics.

An additional and promising method for risk and compliance analysis is an ever-evolving family of outlier or anomaly detection algorithms doing exactly what their name suggests. It is possible and often even preferable to assess and prioritize risks

before they become an issue, such as behavioral deviations representing early signs of money laundering or tax evasion. For this purpose, mining for subtle patterns in a large amount of ongoing transactions or similar structured activity data is needed, which usually means that models cannot be trained based on individual expert-annotated cases of known issues. It is important to note that DSP solutions need to be adjusted to individual financial institutions, even those operating in the same market with the same regulations. Therefore, any case-by-case tuning or expert annotation is impractical.

Unsupervised methods are often used as a first assessment of potential risk and compliance issues. These methods are usually simple, interpretable, and computationally inexpensive, enabling a lot of flexibility and experimentation. Suggestions prepared by these models are usually much cheaper to investigate and validate. If interesting or suspicious ones are detected, more complex models could be appraised to focus on the selected set of cases. By switching the probabilities, a cooperative solution is very likely. The performance of unsupervised methods can be assessed according to the number of experts-approved red flags, allowing straightforward risk calibration.

7.6. Conclusion

Machine learning has been investigated in risk and compliance to see how it might assist regulators and banks manage risk and simplify compliance processes. The risk analytics area, dominated by more fundamental predictive analyses, changes as banks explore black-box algorithms like neural networks while striving to explain decisions. Also, an unexpected outcome is that algorithms earlier considered too abstract to apply to risk are being repurposed for compliance analytics. Analytical techniques that best fit in terms of textual analytics and model performance for inspecting money laundering alerts are not those most loved by compliance professionals. Finally, randomly choosing a technology results in unsuccessful projects. For analytics to drive change, bank and vendor teams must balance privacy, ethics, regulation, and explainability and continuously learn from data and unintended exposition effects. The increasing automation of compliance demands new frameworks and architectures to repeatedly inspect, grade, and score data, transaction, and pattern properties. Whether and how machine learning can produce reasonable probabilities is one aspect of many currently unanswered policy and practical questions. Finally, banks investing in machine monitoring need to work on delivering proactive and reactive analytics. The ability to predict a fraud event has enormous ramifications for banks, regulators, and clients. At the same time, it raises moral and political issues related to accountability and management complexity.

Risk and compliance analytics continue to pick up steam while the rest of the change is less immediate. Here it is less clear what a good market exchange model is, when an

exchange could justify the human cost of transition. More ambiguous are the ‘back office’ tasks like augmenting FX payments straight through processing, eliminating DCA emails warning of mismatches, and enforcing agreement cycles for stress and reconciliation reports. Already almost doubly manual and over a long-enforced track record, current models, however, are further removed from the four areas where markets are already partly or fully automating. Below that threshold is it even less clear what a good model is. The understanding is insufficiently precise, and the operational and genetic variance is too large to accurately estimate churn-out ability under scientific approval ranges. Hence, the market for exotic mathematics and generically simple strategic machines is nigh on intractable. As failures, there will be white elephants that feed and aggravate rather than solve the problem. Here, and possibly within the sourced processes, simple approaches that generate good results already might encourage engagement.

7.6.1. Future Trends

However, further progress can be made in various aspects of data analytics in service operations. With the progressive introduction of new technologies, data analytics will inevitably affect every aspect of service operations. New avenues and opportunities arise to produce ontology-based knowledge for better decision-making. All machine learning and statistical model adoption approaches can be improved to account for long-term operational data in production. Given the various factors and challenges encountered, latest technologies and data-driven platforms can be leveraged to explain decisions for better clarity and trustworthiness. As extensive data from various sources are being collected and stored, the interpretation and translation of big data from all these heterogeneous sources into information and knowledge are proving to be a much greater challenge. On the other hand, the data-percolation effect of social networks allows deriving better hidden knowledge from observations and preferences of similar users engaged in social networks than from the initial data they individually collected. For further enhancements of machine learning-based prediction accuracy, to address the challenges of adopting and deploying machine learning models in real-life scenarios, knowledge management techniques and a sound knowledge architecture is needed for interpreting model behaviors and casualties.

In addition to different machine learning techniques on service operational modeling, groundbreaking work is needed on how to successfully integrate model-free, data-driven analytical techniques with the traditional, characteristic-based white-box analytical techniques. With the rapid increase in subjective and human-resource-related service data, it is pertinent to develop parsing-oriented natural language processing techniques and based on them the specific recommendation and advisory systems in lieu of standard

sentiment-oriented modelling. And with the arrival of the self-driving service model, there is a very open question of what type of new operational decisions will be made by machines and how to make them explicable. Advancing government regulation, service operation safety could draw fruitful lessons and insights from the field of game theory, adversarial machine learning and behavior interpretation. Meanwhile, operational decisions could be better visualized through integrated augmented/virtual reality-enhanced understanding tools incorporating data, models and algorithms .

References

Aladebumoye, T. (2025). *Integrating AI-Driven Tax Technology into Business Strategy*. International Journal of Research and Review, 12(1), 224–231.

Gounagias, N. D., Hristu-Varsakelis, D., & Assael, Y. M. (2018). *Using Deep Q-Learning to Understand the Tax Evasion Behavior of Risk-Averse Firms*. arXiv.

Savić, M., Atanasijević, J., Jakovetić, D., & Krejić, N. (2021). *Tax Evasion Risk Management Using a Hybrid Unsupervised Outlier Detection Method*. arXiv.

Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. (2022). *Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research*. arXiv.

Zheng, X., Zhu, M., Li, Q., Chen, C., & Tan, Y. (2018). *FinBrain: When Finance Meets AI 2.0*. arXiv.