International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 12, pp 327-340, December-2022

https://doi.org/10.47001/IRJIET/2022.612062

Data Quality Management for Effective Machine Learning and AI Modelling, Best Practices and Emerging Trends

Praneeth Reddy Amudala Puchakayala

Data Scientist, Regions Bank, USA. E-mail: appraneethreddy01@gmail.com

***Abstract* - In the modern day, the incorporation of artificial intelligence and machine learning in order to fulfil the requirements of user service has resulted in the establishment of a strong association between data quality and application providers. There are several challenges that come up as a result of the processing of huge amounts of data. These challenges include redundant data, unstructured data, data interruptions, discrepancies, inaccuracies, and information that is no longer relevant. The majority of the attention being paid to data defects in invariant scenarios and the discussion of the eight principles associated to data problems are being directed toward the numerous data quality challenges that are now being addressed. In order to address the issues associated with data quality, a variety of approaches are utilized, which therefore makes it easier to include machine learning and artificial intelligence. It is possible to successfully utilize dataset values in pairs within machine learning models. This is done in order to boost the relevance of the machine learning process through the utilization of a variety of approaches. The process of machine learning involves recognizing patterns and utilizing previous data to generate predictions or decisions. A number of repercussions were investigated on a different level, but the quality of the data was ignored, which resulted in the AI system's trustworthiness and effectiveness being undermined. After everything is said and done, a multitude of real-time applications are investigated for large-scale data in order to guarantee the stability of the data by resolving many risks and concerns regarding privacy. Evaluating a wide range of performance measures ensures that data quality is maintained alongside the integration of AI and ML.**

Keywords: Data Management, Machine learning, Data extraction, IT organization, Transformation, Data quality.

I. INTRODUCTION

The emulation of human intelligence in machines, referred to as artificial intelligence (AI), is extensively utilised across multiple fields, and the volume of scholarly publications in this domain is markedly rising [1]. Artificial Intelligence (AI) refers to the capability of machines to execute tasks that emulate cognitive functions of the human mind, including learning, reasoning, and problem-solving. Machine learning (ML) is a subset of artificial intelligence (AI) characterised by the development of computer algorithms that utilise data to identify patterns, provide predictions, and enhance their performance through further data. Most machine learning methods necessitate substantial quantities of labelled data, leading to a strong collaboration between machine learning and customer science initiatives [4,5]. Customer science, defined as public involvement in scientific research, has expanded markedly in recent years due to technology innovations, including enhanced smartphone capabilities and widespread high-speed Internet access globally. The expansion of customer research has led to extensive dataset collections across several scientific fields, which can serve as a great resource for machine learning algorithms [6] and [7].

While the integration of machine learning and customer science is not novel [8], both domains have predominantly been applied independently until recently [9]. The amalgamation of machine learning with customer science can yield a novel educational framework for customer scientists via human-computer interactions [10]. Furthermore, it may lead to enhanced multidisciplinary cooperation among researchers and the public across several domains, including computer science, ecology, astronomy, and medicine, among others [9]. This integration has mostly concentrated on object detection in photos and videos, specifically targeting automatic species identification in biodiversity initiatives [11]. The iNaturalist project [12] is a prominent example that has incorporated automatic species identification suggestions since 2017, utilising photographs provided by observers. The automatic identification has enhanced over the years due to the increased utilisation of photos for model training, with the most recent model release occurring in March 2020 at the time of this writing. The automated species identification in iNaturalist has afforded customer scientists the chance to acquire knowledge about species and to reduce the incidence of inaccurate observations [14].

The aim of integrating customer science with machine learning extends beyond only supplying data for algorithms and automating identification tasks. The objective is to integrate human and machine intelligence to enhance customer scientific activities, including automated data collecting,

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 6, Issue 12, pp 327-340, December-2022
https://doi.org/10.47001/IRJIET/2022.612062

processing, and validation, while also augmenting public participation.

Data is considered one of the most valuable assets for managing automation systems in the period of technological innovation, when several types of data are available due to advancements in information technology. Consequently, financial markets and technological advancement have been intricately connected to all human activities during the past several decades. Big data technology has emerged as a crucial element of the financial services sector and will propel future innovation. Financial innovations are regarded as the most rapidly evolving topic within financial services. They encompass a variety of financial firms, including online peer-to-peer lending, crowdfunding platforms, SME financing, wealth management, asset management platforms, trading management, cryptocurrency, money and remittance transfer, and mobile payment systems, among others. Each of these services creates thousands of data items daily. Consequently, in this era of technological innovation, various forms of data are accessible due to advancements in information technology; data is regarded as one of the most valuable assets in the administration of automated systems [3]. Overseeing this data is regarded as the key element of these services. Any data loss may lead to substantial complications for the specific financial sector. Financial analysts now utilise external and alternative data to inform their investing decisions.

Moreover, financial industries utilise big data to develop sophisticated decision-making models by extensive predictive studies and monitoring various spending trends. Industries can select the financial products they wish to offer in this manner [4]. Financial institutions exchange millions of data points. Consequently, big data is garnering more attention within the financial services sector, where information profoundly influences critical production and success factors. It has become increasingly essential to enhance our comprehension of financial markets, and the financial sector consistently employs trillions of data points in daily decision-making. The future of the financial services business is critically dependent on advancements in trade and investment, tax reform, fraud investigation and detection, risk analysis, and automation. Furthermore, it has revolutionised the financial sector by overcoming challenges and acquiring valuable insights to improve client satisfaction and the overall banking experience [6]. Big data is transforming finance in five ways: transparency, risk analysis, algorithmic trading, consumer data utilisation, and cultural shift. Moreover, big data significantly influences economic analysis and economic models.

Acquired data is invariably imperfect. Regardless of the sophistication of data quality tools and automated solutions employed, technology alone cannot comprehensively address your organization's data quality issues. Businesses have undoubtedly encountered the following data quality challenges in data pipelines at least once. Let us examine the eight principal data concerns and explore how specialists address data quality challenges.

This study has compiled and analysed the perspectives of numerous academics, researchers, and others on big data and financial activity. This study aims to assess the current theory while also gaining a profound understanding of the research via qualitative data. Nonetheless, the investigation of big data within financial services is less extensive compared to other financial industries. Few works specifically examine big data in diverse financial research contexts. Comprehensive analyses of big data in financial services remain scarce, despite several research addressing specific subjects. Consequently, the necessity to identify the financial sectors significantly influenced by big data is fulfilled. The investigation of big data and financial difficulties is very contemporary.

In order to guarantee that the effectiveness of the relationship between the data quality and the application is enhanced through the integration of AI and ML. Obtaining the data is necessarily going to be an imperfect issue, and it will be detrimental to the operations of the organization. There are a variety of challenges that are depending on the quality of the data. When it comes to entirely resolving difficulties with data quality, there is no solution that is more effective than training and raising the awareness and professional abilities of the staff working for the organization. Data that has been acquired is inevitably error-prone. Regardless of the level of sophistication of the data quality tools and automated solutions that are utilized, technology cannot, on its own, provide a full solution to the data quality problems that your organization is experiencing. There is little doubt that there have been instances in which businesses have been confronted with the following data quality difficulties in data pipelines. Let us take a look at the eight most important data problems and investigate the ways in which specialists manage the challenges associated with data quality, as was covered in section 2.

The improvement of the data quality must be ensured based on various elements, such as the precision of the data, the consistency of the data, the completeness of the data, the recentness of the data, and the relevance of the data. There is a widespread belief among numerous organizations that the collection and storage of comprehensive customer data will, at some point in the future, result in advantageous outcomes. All the same, that is not necessarily the case in every circumstance. It is possible for organizations to face the

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 12, pp 327-340, December-2022

https://doi.org/10.47001/IRJIET/2022.612062

difficulty of irrelevant data quality as a result of the huge number of data, not all of which is immediately relevant.

In conclusion, machine learning and artificial intelligence are combined in order to automate and recognize the underlying pattern and their link in order to identify the data patterns and predictions based on the data. Linear models, ensembles, boosted models, and neural networks are some of the machine learning models that are explored in this article.

This study primarily ensures the utilization of customer services being offered with the integration of ML and AI to enable automation for large amount of data with effective quality.

This paper help to ensure the effective relationship between the customer and application providers based on the integration of AI with the customer requirement. As the large amount of data are carried out, there are various challenges and issues are addressed in terms of redundant data, unstructured data, data interruption, data discrepancies, Error in the data and outdated data. Ensuring the improving the parameters such as,

- Data Precision
- Data Consistency
- Completeness
- Data Recency
- Data Relevancy

Finally, to identity the data patterns and predictions based on data, ML and AI is integrated to automate and recognize the underlying pattern and their relationship. Various ML models are discussed as linear models, ensembles, boosted models, and neural networks.

II. MOST COMMON DATA QUALITY ISSUES / CHALLENGES

Acquired data is invariably imperfect. Regardless of the sophistication of data quality tools and automated solutions employed, technology alone cannot comprehensively address your organization's data quality issues. Businesses have undoubtedly encountered the following data quality challenges in data pipelines at least once. Let us examine the eight principal data concerns and explore how specialists address data quality challenges.

A) Redundant Data

Duplicate data denotes a particular system or database that retains several instances of the identical data record or information. Common reasons of data duplication include multiple re-imports, inadequate decoupling in data integration procedures, acquisition from several data sources, and data silos. For example, if an auction item is posted on an auction website twice, it may adversely impact both prospective purchasers and the website's reputation. The presence of duplicate records can result in inefficient storage utilisation and elevate the likelihood of distorted analytical outcomes [15].

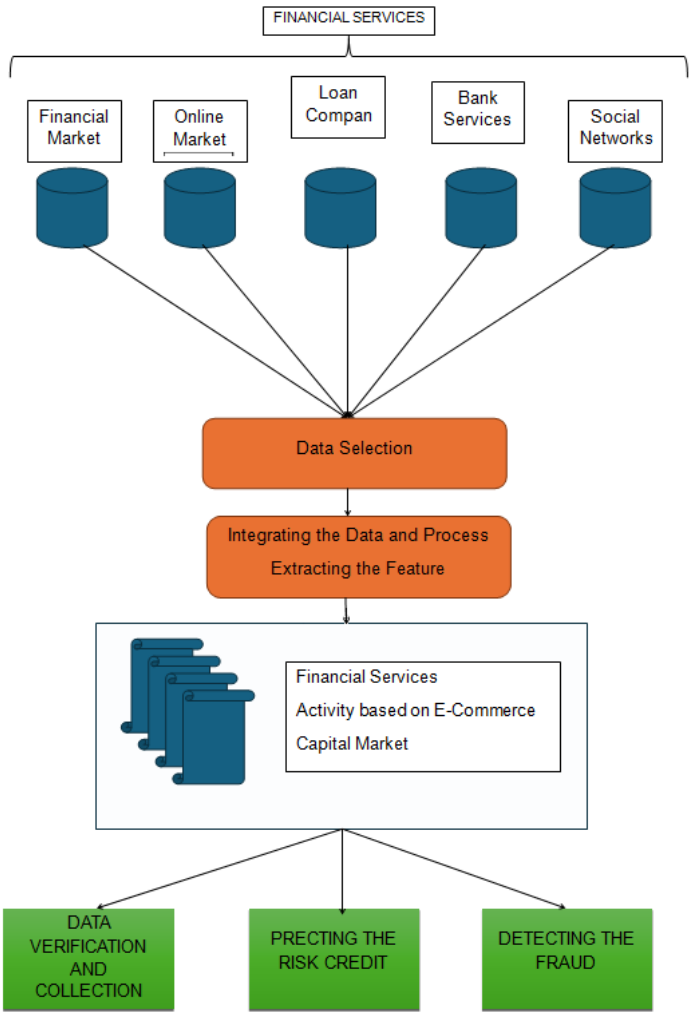

Figure 1: Financial Data Processing

- Implement data governance frameworks that encompass data entry and storage protocols.
- Implement data validation checks prior to inputting data sets into the system.
- Employ distinctive identities for data elements such as customers, goods, and products.
- Utilise data duplication software to detect quality flaws and eliminate them from systems.
- Engage in manual data cleansing operations.
- Inconsequential Information.

Numerous organisations contend that the acquisition and retention of comprehensive consumer data will yield advantages at some future juncture. Nonetheless, that is not inherently the situation. Due to the vast volume of data, not all

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 12, pp 327-340, December-2022

https://doi.org/10.47001/IRJIET/2022.612062

of which is instantly pertinent, businesses may encounter the challenge of irrelevant data quality[16]. Prolonged storage of unnecessary data leads to rapid obsolescence and diminishes its value, while encumbering IT infrastructure and absorbing the management resources of data teams.

- Specify the data needs, including data pieces, sources, and other pertinent details, for a project.
- Employ filters to eliminate extraneous data from extensive data collections.
- Identify and utilise the appropriate data resources pertinent to the project.
- Employ data visualisation to emphasise pertinent patterns.

B) Unstructured Data

Unstructured data may be seen as a data quality concern due to many issues. Unstructured data denotes any type that lacks a specific data structure or model, including text, audio, and images, making it difficult for organisations to store and analyse effectively [17].Similar to other forms of raw data, unstructured data originates from various sources and may contain duplicates, irrelevant content, or erroneous information. Transforming unstructured data into valuable insights is challenging, necessitating specialised tools and integration methodologies. This issue now pertains to skill and the recruitment of data analysts rather than cost as in figure 2.

Businesses may prioritise structured data over unstructured data when they lack the requisite talents and resources to manage the latter effectively. Prior to eliminating unstructured data assets from the database, meticulously assess the disparity between the investment expenditures and the latent advantages.

- Utilise automation and technology such as artificial intelligence (AI), machine learning (ML), and natural language processing (NLP).
- Recruit and educate individuals possessing expertise in data administration and analysis.
- Formulate data governance policies to direct data management activities within the organisation.
- Implement data validation tests to restrict the input of unstructured data.

C) Data Interruption

Data downtime denotes the interval during which data is either unprepared or entirely unreachable. Data outage results in organisations and customers losing access to essential information. This unintentionally undermines the requirements of these audiences, resulting in suboptimal analytical outcomes and customer grievances [18].

Common variables contributing to data downtime may include unexpected schema alterations, migration complications, technical malfunctions, and network or server failures, all influenced by the condition of the management system. To restore data to the system and prevent data outages, a data engineer must dedicate time to upgrading and ensuring the integrity of the data pipelines. Prolonged data maintenance and storage incurs greater resource expenditure for the organisation, adversely impacting customer trust.

- Establish redundancy and failover systems, including backup servers and load balancing, to guarantee the continuous availability of vital data.
- Perform routine maintenance and updates prior to data unavailability.
- Assess data pipeline efficiency and network bandwidth to detect potential problems.
- Automate the data management process by instituting validation and verification procedures, among others.

D) Discrepant data

Mismatches in identical information across many sources are unavoidable due to the diverse origins of data [19]. This phenomenon is collectively referred to as "inconsistent data." Data discrepancies occur due to various circumstances, including manual entry errors and ineffective data management techniques. One thing you may not have considered is the disparity in units and language. The representation of a date serves as a pertinent example. A date can be represented in several formats according to the source's needs, such as April 14, 2023; 14/04/2023; or 04-14-2023. In this instance, there is no incorrect date format. Nonetheless, it significantly impacts the quality of the data. Inconsistent data, irrespective of its cause or format, deteriorates data integrity and undermines the intrinsic value of data, disrupting all business operations.

- Implement data governance principles to ensure consistency in formatting across sources.
- Utilise technologies such as artificial intelligence (AI), machine learning (ML), and natural language processing (NLP) to automate the identification and rectification of contradictory data.
- Consistently audit and sanitise data systems.
- Streamline the data entering procedure by employing drop-down menus or data picklists.

E) Erroneous data

Inaccurate data refers to information that contains inaccuracies, compromising its quality and trustworthiness. Due to its expansive nature, various data quality concerns, including incompleteness, obsolescence, inconsistency,

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 12, pp 327-340, December-2022

https://doi.org/10.47001/IRJIET/2022.612062

typographical errors, and absent or erroneous numbers, are also classified as faulty data. For actionable insights to be generated, it is essential that the acquired data be very accurate [20] and accurately represents reality. Nonetheless, due to many external and internal causes, including human error during data input and data drift, the information cannot maintain its original accuracy as it was at the time of entry. Inaccurate data leads to erroneous company decisions and results in customer dissatisfaction with the information presented.

Inaccurate customer information within a CRM database results in ineffective marketing campaigns, diminished income, and consumer discontent.

- Formulate data governance principles by directing data quality requirements.
- Employ data cleansing methodologies, including data normalisation, to rectify inaccuracies.
- Automate data quality procedures via data profiling technologies and data validation frameworks.
- Regularly review and clean the data system.

F) Concealed information

Organisations extract and analyse data to enhance operational efficiency. Nevertheless, given the vast volume of data available today, most organisations utilise only a fraction of it. The residual underused or absent data within data silos is termed hidden data. More specifically, hidden data can be valuable but unused and stored within other files or documents or invisible information to customers, such as metadata [21].

For example, a company's sales team possesses data about clients, yet the customer service staff lacks this information. The organisation may forfeit the chance to develop more precise and comprehensive consumer profiles without disclosing the requisite information. Concealed data should be utilised or eradicated. Having this data quality issue present is not only a way to waste resources; hidden data can even result in privacy monitoring and compliance violations if they are sensitive data within datasets.

- Invest in data catalog solutions.
- Use data masking to replace sensitive data with fictitious data while retaining the original data format.
- Use machine learning algorithms to identify hidden data.
- Limit access to certain data types based on employee roles and responsibilities.

G) Outdated Data

Collected data can become obsolete quickly and inevitably lead to data decay through the development and modernization of human life [22]. All information that is no longer accurate or relevant in the current state is considered outdated data. Information about a customer, such as name, address, contact details, etc., is a good example that needs to be constantly updated so as not to miss opportunities to consult about the company's services and promotions.

The problem of old data is not only a concern about accuracy, but it also reflects the delay and lack of investment and interest of enterprises in database management systems. The consequences of outdated data can extend to incorrect insights, poor decision-making, and misleading results.

- Regularly review and update data.
- Establish a data governance strategy to effectively manage data.
- Use outsourcing services in data management if managing data in-house is not feasible.
- Use machine learning algorithms to identify outdated data.

Regardless of the type, data quality issues all harm business operations. Orient Software has provided you with valuable expert solutions to limit or even eliminate such problems through this article. However, human is the core value of every organization. There is no better solution to completely solve data quality issues than training and improving the awareness and professional skills of the company's personnel.

When you deploy a machine learning model in production, it faces real-world data. As the environment changes, this data might differ from what the model has seen during training. As a result, the so-called data drift can make the model less accurate [23].

Data drift refers to changes in the distribution of the features an ML model receives in production, potentially causing a decline in model performance. When ground truth labels aren't accessible, data drift monitoring techniques serve as proxy signals to assess whether an ML system operates under familiar conditions. Data drift is a change in the statistical properties and characteristics of the input data. It occurs when a machine learning model is in production, as the data it encounters deviates from the data the model was initially trained on or earlier production data.

This shift in input data distribution can lead to a decline in the model's performance. The reason is, when you create a machine learning model, you can expect it to perform well on data similar to the data used to train it. However, it might struggle to make accurate predictions or decisions if the data keeps changing and the model cannot generalize beyond what it has seen in training.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 12, pp 327-340, December-2022

https://doi.org/10.47001/IRJIET/2022.612062

In simple terms, data drift is a change in the model inputs the model is not trained to handle. Detecting and addressing data drift is vital to maintaining ML model reliability in dynamic settings.

III. METHODS TO OVERCOME DATA QUALTIY ISSUES

This concept is especially pertinent in the realm of AI, as AI models, encompassing machine learning and deep learning, are fundamentally dependent on the data utilised for training and validation. The AI model is likely to yield unreliable or biassed outcomes if the training data is biassed, incomplete, or erroneous. To prevent the GIGO issue, it is essential to guarantee that the data utilised in AI systems is precise, representative, and of superior quality. This typically entails data cleansing, preprocessing, and augmentation, alongside the application of rigorous assessment measures to evaluate the efficacy of AI models [24].

- Precision:
  Accurate data is essential for AI systems, allowing them to generate correct and dependable results. Data entry errors can result in erroneous judgements or misleading insights, thereby harming organisations and individuals.
- Consistency:
  Guarantees that data adheres to a uniform format and structure, hence enhancing the efficiency of data processing and analysis. Inconsistent data can result in confusion and misinterpretation, undermining the efficacy of AI systems.
- Completeness:
  Incomplete datasets might result in AI algorithms overlooking critical patterns and connections, hence producing incomplete or biassed outcomes. Guaranteeing data completeness is essential for the accurate and comprehensive training of AI models.
- The recency of data significantly influences AI performance.
  Obsolete data may fail to represent the contemporary environment or trends, leading to inappropriate or deceptive outputs.
- Relevance:
  Pertinent data immediately addresses the issue at hand, enabling AI systems to concentrate on the most significant factors and relationships. Extraneous data can obfuscate models and result in inefficiencies.

A) In machine learning and AI development, imputation techniques are essential for addressing data quality challenges such as missing values by employing algorithms to estimate and substitute gaps with predicted values derived from patterns identified in the existing data. This enables models to train on a more comprehensive dataset, yielding more precise outcomes, particularly when managing incomplete or partially missing information in real-world datasets.

Essential aspects of imputation techniques and data quality challenges in machine learning and artificial intelligence. Absence of values in a dataset can considerably impair the efficacy of machine learning models, resulting in skewed predictions and incorrect outcomes [25].

The purpose of imputation is to estimate absent values by utilising the correlations among existent variables in the dataset, thereby effectively "filling in" the voids with credible values.

- Mean/Median Imputation:
  Substituting absent values with the mean or median of the relevant characteristic, a straightforward method that may lack precision if the data deviates from a normal distribution.
- K-Nearest Neighbours (KNN) Imputation: Determining analogous data points through other attributes and utilising their values to fill in missing data.
- Multiple Imputation (MI): Generating several iterations of the dataset with varied imputed values to more effectively represent the uncertainty linked to absent data.
- Matrix Factorization-based Imputation: Employing matrix factorisation methods to discern latent patterns within the data and estimate missing values according to these patterns.
- Complex algorithms:
  Machine learning models such as decision trees, neural networks, and regression algorithms may elucidate intricate relationships within data and forecast absent variables with enhanced precision.
- Adaptive Imputation:
  Machine learning models can modify imputation algorithms according to the specific context of absent data, taking into account elements such as the mechanism of missingness and interactions among features.
  - Enhanced model performance: By rectifying missing data, imputation approaches can substantially improve the precision and dependability of machine learning models.
  - Management of incomplete datasets: Facilitates the use of datasets that may contain absent values, a frequent occurrence in practical situations.

B) Normalisation and standardisation techniques in machine learning and AI development are essential data preprocessing steps that address data quality issues by ensuring features are

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 12, pp 327-340, December-2022

https://doi.org/10.47001/IRJIET/2022.612062

on a comparable scale, thereby facilitating more accurate model training and analysis, particularly when handling datasets with significantly disparate variable ranges. Essential aspects of normalisation and standardisation in artificial intelligence and machine learning:

Rectifying disparate scales:

When features in a dataset exhibit markedly disparate ranges (e.g., income versus age), algorithms may exhibit bias towards features with greater magnitudes. Normalisation and standardisation facilitate uniformity by adjusting all features to a comparable range [26].

Enhancing model convergence: By approximating data to a normal distribution, these methods help expedite and stabilise the training of machine learning models, particularly those dependent on distance metrics such as K-Nearest Neighbours (KNN) [27].

Prevalent normalisation and standardisation methodologies:

- Min-Max scaling (Normalisation): Adjusts values to a range between 0 and 1 by subtracting the minimum value and dividing by the feature's range.
- Z-score standardisation: Adjusts data to achieve a mean of 0 and a standard deviation of 1 by removing the mean and dividing by the standard deviation.
- Utilising AI and ML to tackle data quality challenges beyond mere normalisation and standardization.
- Anomaly detection: Machine learning algorithms can discern outliers and atypical data points that may signify flaws or inconsistencies within the data.
- Missing data imputation involves employing algorithms to populate absent values by analysing patterns within the available data.

Data cleansing and verification:

- Automated verifications to detect and rectify erroneous data entries, such as improper formatting or conflicting values.
- Feature engineering involves generating additional features from existing data to enhance model performance and mitigate any data quality concerns.

The significance of data quality for artificial intelligence and machine learning.

- Model precision: Substandard data results in erroneous predictions and incorrect model outputs.
- Bias mitigation: Rectifying data biases via cleansing and normalisation helps avert models from acquiring distorted patterns.
- Generalisability: High-quality data guarantees the model's efficacy on novel, unobserved data.

C) In machine learning and AI development, SMOTE (Synthetic Minority Oversampling Technique) and under sampling methods are predominantly employed to rectify data quality concerns associated with class imbalance, wherein one category possesses markedly fewer samples than others. These techniques artificially generate supplementary data points for the minority class, thereby "balancing" the dataset and enhancing the model's capacity to learn from the under-represented class [28].

Essential aspects of SMOTE and under sampling:

- SMOTE: Generates synthetic data points for the minority class by examining the attributes of current minority samples and producing new ones inside the feature space, thereby "bridging the gaps" between existing data points.
- Under sampling: Decreases the quantity of samples from the majority class to attain a more equitable distribution.

Advantages of employing SMOTE and under sampling:

- Enhanced model efficacy:
- By rectifying class imbalance, these methods can markedly improve the precision of machine learning models, particularly in forecasting occurrences from the minority class.
- Enhanced generalisation:
- SMOTE enhances the model's ability to learn robust patterns and generalise to unknown data by creating synthetic data points.

Appropriate application of each technique:

- SMOTE: When the minority class is markedly under-represented, and there is a desire to retain maximal information from the current minority samples. When sufficient processing power is available to produce synthetic data points.
- Under sampling: When the majority class is excessively large, eliminating certain samples will not lead to a substantial loss of information. When addressing high-dimensional data, the generation of synthetic samples may prove computationally prohibitive.

D) In machine learning and AI development, outlier detection approaches are essential for mitigating data quality issues by finding and rectifying anomalous data points that might substantially compromise model accuracy, hence facilitating cleaner and more dependable data for training and analysis [29].

ISSN (online): 2581-3048
Volume 6, Issue 12, pp 327-340, December-2022
https://doi.org/10.47001/IRJIET/2022.612062

Essential aspects of employing outlier identification to address data quality challenges:

- The presence of outliers can substantially distort the outcomes of machine learning models, resulting in erroneous predictions and deceptive inferences if inadequately addressed.
- The function of machine learning involves the proficient identification of outliers through the analysis of data patterns, pinpointing data points that substantially diverge from the anticipated distribution.
- Prevalent Outlier Detection Methods:

Statistical Techniques:

- Z-score: Computes the standard deviation of data points and detects outliers according to a threshold, usually set at 2 or 3 standard deviations.
- Interquartile Range (IQR): Determines outliers by assessing the interval between the first and third quartiles.
- Density-Based Approaches:
- Local Outlier Factor (LOF): Assesses the density of data points surrounding a certain site, identifying outliers as points exhibiting markedly low density.
- Algorithms for Anomaly Detection:
- Isolation Forest: Randomly divides the data space, resulting in outliers being segregated into smaller segments.
- One-Class SVM (OCSVM): Develops a model based on "normal" data and detects outliers as instances that reside beyond the established limit.

Data Preparation:

- Employ outlier identification methodologies during data cleansing to discover and manage outliers prior to model training.
- Feature Engineering: Develop novel features explicitly intended to emphasise potential outliers.
- Model Surveillance: Consistently assess model efficacy to detect novel anomalies in real-time data streams.

E) The development of machine learning (ML) and artificial intelligence (AI), in conjunction with drift detection algorithms, can effectively mitigate data quality challenges by proactively recognising and rectifying alterations in data distribution over time. This ensures that models maintain accuracy and reliability as the underlying data evolves, accomplished through continuous monitoring and necessary model adaptation, thereby preventing "data drift" that could result in suboptimal predictions if neglected [30].

Essential elements of how machine learning and artificial intelligence can mitigate data quality challenges through drift detection:

- Detecting Data Anomalies: Machine learning algorithms can scrutinise extensive datasets to identify outliers, inconsistencies, and trends that may signify data quality issues, facilitating precise data cleansing and rectification.
- Real-time Monitoring: By employing drift detection techniques such as the Drift Detection Method (DDM) or the Page-Hinkley test, one can perpetually oversee data streams for alterations in distribution, activating alerts upon the occurrence of substantial variations.
- Adaptive Model Retraining: Upon detecting drift, machine learning models can be autonomously retrained with new data to adjust to the evolving data distribution, hence maintaining accuracy over time.
- Feature Engineering: By meticulously selecting and modifying features, one can reduce model sensitivity to potential data shifts, hence augmenting robustness against drift.

Prevalent drift detection algorithms:

- Statistical Analyses:
- The Kolmogorov-Smirnov test evaluates the cumulative distribution functions of various data sets.
- Chi-Square test: Evaluates categorical data for substantial discrepancies in distributions.
- Adaptive Thresholding Algorithms: Page-Hinkley Test: Continuously observes data and triggers an alert upon the detection of a statistically significant change.
- Drift Detection Method (DDM): Employs a sliding window to juxtapose current data against previous data, identifying drift through statistical criteria.

Advantages of employing machine learning and drift detection for data quality.

- Enhanced Model Performance: Proactive identification and alleviation of data drift result in more precise and dependable predictions from your machine learning models.
- Minimised Manual Intervention: Automated data quality assessments and drift detection substantially diminish the necessity for manual data cleansing and validation.
- Early Warning System: By detecting potential data quality concerns promptly, corrective measures can be implemented prior to their substantial impact on your models.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 12, pp 327-340, December-2022

https://doi.org/10.47001/IRJIET/2022.612062

IV. INCORPORATE DATA QUALITY CHECKS IN THE PRODUCTION PIPELINE

Machine learning models are algorithms designed to identify patterns and make predictions or decisions based on data. These models are trained using historical data to recognize underlying patterns and relationships. Once trained, they can be used to make predictions on new, unseen data. Modern businesses are embracing machine learning (ML) models to gain a competitive edge. It enables them to personalize customer experience, detect fraud, predict equipment failures, and automate tasks [31].

Hence, improving the overall efficiency of the business and allow them to make data-driven decisions. Deploying ML models in their day-to-day processes allows businesses to adopt and integrate AI-powered solutions into their businesses. Since the impact and use of AI are growing drastically, it makes ML models a crucial element for modern businesses. In the context of machine learning, model testing refers to a detailed process to ensure that it is robust, reliable, and free from biases. Each component of an ML model is verified, the integrity of data is checked, and the interaction among components is tested. The main objective of model testing is to identify and fix flaws or vulnerabilities in the ML system.

It aims to ensure that the model can handle unexpected inputs, mitigate biases, and remain consistent and robust in various scenarios, including real-world applications. Since testing ML models is a very important task, it requires a thorough and efficient approach. Multiple frameworks in the market offer pre-built tools, enforce structured testing, provide diverse testing functionalities, and promote reproducibility. It results in faster and more reliable testing for robust models.

1. For analyzing how your model will behave in the real world, you must split your dataset into training, testing, and validation data. The training data set can be used for model training, the testing data set can be used for testing on unknown data, and the validation data set can be used for hyperparameter tuning and model selection. We may measure the actual model performance by creating multiple samples and splits in the dataset. The number of samples and the model used determine the dataset split ratio. Duplicate samples can emerge in both training and testing sets for various reasons [32].

Therefore it's critical to spot and eliminate them. Before splitting the data, it is best to remove duplicates, check for partial copies, sort by different columns, and review the resulting data. It's critical to create a dataset that closely resembles real-world data and use it to evaluate your model. This is especially significant when the dataset and the production data are not from the same source. To perform this validation, compare the structure of a real-world data point to the structure of your training data. If your model was trained on a clean dataset, it's critical to create a dataset that closely resembles real-world data and use it to evaluate your model. This is especially significant when the dataset and the production data are not from the same source.

To perform this validation, compare the structure of a real-world data point to the structure of your training data. One of the most common causes of model accuracy degradation over time is data drift. It essentially means that the data distribution changes over time and differs from the training data distribution. So, to determine drift, you can train your model on past data and then evaluate it on current data; if the results deviate significantly from the historical data, you're dealing with data drift.

2. By leveraging outlier detection algorithms, pipeline operators can identify outliers, ensuring the accuracy and integrity of the data processed. These methods can aid in the direct classification of pipe conditions without the need for specialized intervention, thereby streamlining the detection process [33]. The use of machine learning algorithms in outlier detection contributes to the identification of anomalies in pipeline systems, allowing for prompt follow-up actions by human experts. Outlier detection methods have been shown to improve the performance of machine learning models, contributing to more accurate and reliable decision-making processes. Automated outlier detection approaches have been developed to replace manual methods, offering a more efficient and less error-prone solution for identifying outliers in real data. Early detection of outliers in wireless sensor networks has been highlighted as a means to conserve network resources and prevent unnecessary transmissions.

Outlier detection is a critical component in various industrial fields including data science. Machine learning techniques are commonly used for outlier detection due to their effectiveness in identifying anomalies in raw data. Machine learning models trained on diverse datasets have shown improved accuracy and efficiency in detecting outliers. The combination of supervised and unsupervised techniques in machine learning frameworks has demonstrated the potential to enhance outlier detection effectiveness. Furthermore, ensemble algorithms like XGBOD have been introduced to improve the detection of outliers in various datasets. These advanced algorithms and techniques contribute to the enhanced identification of anomalies within datasets, thereby aiding in outlier detection processes. Moreover, outlier detection methods have found applications in various domains, including healthcare and IoT, underscoring their versatility and effectiveness in identifying anomalies.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 12, pp 327-340, December-2022

https://doi.org/10.47001/IRJIET/2022.612062

The utilization of outlier detection techniques across diverse fields highlights their importance in addressing critical issues such as risk assessment, anomaly identification, and complex disease characterization [34]. The integration of acoustic emission data with machine learning algorithms is crucial for achieving rapid and accurate leak detection, emphasizing the importance of real-world data integration with artificial intelligence. Novel algorithms combining statistical methods and machine learning clustering techniques have been proposed to enhance outlier detection accuracy and control extreme values more effectively. Machine learning-based anomaly detection methods offer flexibility, stability, and reliability, with a low probability of false alarms, ensuring robust outlier detection capabilities. In practice, once ILI data is collected, it requires further analysis to understand pipeline's existing condition, identify anomalies, corrosion growth, and other defects.

To achieve these objectives, a detailed analysis is required to understand the data, remove outliers, and make decisions that help pipeline owners in replacing or repairing the pipeline or its particular segment. Inaccuracies in measurement, calibration, or recording might result in incorrect assessments of the reliability of the pipeline. Inaccurate data can lead to incorrect maintenance decisions, which may result in unnecessary expenses or, more importantly, missing real threats to pipeline integrity.

3. Data drift refers to changes in the distribution of the features an ML model receives in production, potentially causing a decline in model performance. When ground truth labels aren't accessible, data drift monitoring techniques serve as proxy signals to assess whether an ML system operates under familiar conditions. You can use various approaches to detect data distribution drift, including monitoring summary feature statistics, statistical hypothesis testing, or distance metrics.

Data drift is a change in the statistical properties and characteristics of the input data. It occurs when a machine learning model is in production, as the data it encounters deviates from the data the model was initially trained on or earlier production data. This shift in input data distribution can lead to a decline in the model's performance [35]. The reason is, when you create a machine learning model, you can expect it to perform well on data similar to the data used to train it. However, it might struggle to make accurate predictions or decisions if the data keeps changing and the model cannot generalize beyond what it has seen in training. In simple terms, data drift is a change in the model inputs the model is not trained to handle. Detecting and addressing data drift is vital to maintaining ML model reliability in dynamic settings.

4. For the machine learning practitioner, data quantity and quality are equally critical. Given a restricted budget, practitioners must balance the costs of getting more data against the costs of acquiring better data or cleaning data. It is occasionally feasible to acquire valuable insights from low-quality data, but low quality can also lead to wholly incorrect and biassed models, particularly if the data is systematically biassed rather than simply noisy.Several data-quality models list quality criteria, such as how data is stored and safeguarded.

For example, ISO/IEC standard 25012 has 15 criteria: accessibility, accuracy, availability, completeness, compliance, secrecy, consistency, credibility, currentness, efficiency, portability, precision, recoverability, and traceability.

Data quality issues might arise for any criterion. For example, in our inventory system, someone could enter the wrong product or number of items when a shipment is received (accuracy), simply forget an item or the entire shipment (completeness), enter the shipment twice (uniqueness), enter different prices for the same item in two different tables (consistency), or simply forget to enter the data until a day later (currentness). Depending on the system, quality issues for some criteria may be more significant than others.

V. CONSEQUENCES OF IGNORING DATA QUALITY

When developing machine learning and AI systems, ignoring data quality can lead to significant risks, including biassed outputs, inaccurate predictions, and flawed decision-making, ultimately damaging the credibility and effectiveness of the AI system, potentially causing financial losses and reputational harm to the organisation deploying it; essentially, poor data quality can amplify existing biases, generate incorrect results, and open vulnerabilities to manipulation [36].

Our primary goal was to investigate temporal model degradation trends observed in the most prominent machine learning algorithms utilised in ML/AI projects. As a result, we have chosen the following four major model types, incorporating four primary approaches to ML model creation (linear models, ensembles, boosted models, and neural networks).

1. RidgeRegressor model is penalised linear regression.
2. Random Forest Regressor Model (RF).
3. Gradient boosting in the XGBoost model (XG).
4. Neural network MLPerceptronRegression model (NN).

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 12, pp 327-340, December-2022

https://doi.org/10.47001/IRJIET/2022.612062

We purposefully chose these model types to represent classical yet fundamentally diverse mathematical approaches to ML model building and training, as well as to incorporate well-established optimisation and regularisation techniques that provide stability in the face of unknown data, noise, and overfitting. By analysing ageing trends across several model types, we were able to identify parallels and variations in how different models age on the same data.

To eliminate domain bias, we then chose 32 datasets from four industries, representing completely distinct processes and target variables.

1. Weather prediction for the next day's temperature and humidity (city of Basel, 2010-2020 data)29.
2. Predicting patient examination delays in four outpatient facilities in a major regional hospital (30).
3. Airlines predict aeroplane departure delays at 15 domestic US airports31.
4. Financial Predicting the Next Day's Stock Closing Value (11 S&P500 stocks).

We ensured that our analysis only included datasets with

- Multiple years of timestamped data records,
- No missing/partial data, and
- Several variables not directly derived from time (to account for the indirect impact of temporal patterns).

We also guaranteed that each model-dataset pair provided a satisfactory prediction quality, with cross-validated R2 values ranging from 0.7 to 0.9 throughout model training. As a result, we were able to focus our efforts just on data/models with high initial quality, as is typical throughout model deployment in practice.

Finally, we verified that none of the preceding datasets contained any rapid changes in the target variable's value. It is reasonable to predict a large loss in model quality when the underlying data changes rapidly. However, it is significantly more disturbing to notice model quality erosion when the data remains consistent, with nothing alerting users to potential problems [37].

In machine learning and AI development, a significant potential risk is "bias in the model," which occurs when an AI system produces skewed or discriminatory results due to biases present in the data used to train it, resulting in unfair outcomes for certain groups of people; this bias can stem from initial data collection, algorithm design, or even societal prejudices reflected in the training data.

Key points about bias in AI models:

- Bias originates from:
- Bias can be introduced through several means, including:
- Biassed training data: If the data used to train the AI model is not representative of the population it is supposed to serve, the model will learn and reproduce such biases.
- Algorithmic design: The algorithms used and how they are implemented may inadvertently favour certain outcomes over others.
- Human biases: Developers' own biases might influence data selection and model construction.

Example of biassed AI outcomes:

- Facial recognition algorithms misidentify persons of colour due to datasets that do not adequately reflect varied ethnicities.
- Job recruitment algorithms favour male candidates if the training data includes more male applications.
- Credit scoring models refuse loans to particular demographics based on previous data that shows discriminatory practices.

Bias has the following potential consequences:

- Discrimination: Artificial intelligence systems can propagate societal biases, resulting in unfair treatment of individuals or groups.
- Loss of faith: Biassed AI can undermine public trust in technology and its uses.
- Legal issues: Companies that use biassed AI systems may face legal consequences for discriminatory practices.

AI/ML systems have made significant advances in the last decade. Although the construction of a machine capable of understanding or learning any intellectual work that a human being undertakes is not within reach, today's AI systems may perform well on tasks that are well defined and often demand human intelligence. The learning process, a crucial component of most AI systems, is represented by machine learning (ML), which is based on mathematics, statistics, and decision theory. Most recent breakthroughs, such as self-driving cars, digital assistants, and facial recognition, can be attributed to advances in machine learning, particularly deep learning algorithms [38].

The financial sector, spearheaded by fintech businesses, is quickly growing its usage of AI/ML technologies (Box 1). The financial sector's recent acceptance of technology developments such as big data and cloud computing, combined with the expansion of the digital economy, has enabled the effective deployment of AI/ML systems. According to a recent study of financial institutions (WEF 2020), 77 percent of respondents expect AI to be of high or

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 12, pp 327-340, December-2022

https://doi.org/10.47001/IRJIET/2022.612062

very high overall importance to their businesses within the next two years. McKinsey (2020a) predicts that the potential value of AI in the banking sector is $1 trillion.

AI/ML systems are reshaping client experiences such as communication with financial service providers (for example, chat bots), investing (for example, robo-advisors), borrowing (for example, automated mortgage underwriting), and identity verification (for example, image recognition). They are also altering financial institution operations, saving significant costs by automating procedures, employing predictive analytics to improve product offerings, and delivering more effective risk and fraud management processes as well as regulatory compliance. Finally, AI/ML systems give central banks and prudential oversight bodies additional tools for monitoring systemic risks and strengthening prudential control.

According to the Bank of England (2020) and McKinsey (2020b), a significant proportion of financial organisations anticipate that AI/ML will play a larger role following the pandemic. Customer relationships and risk management are key growth areas. Banks are looking into methods to use their experience utilising AI/ML to handle the large volume of loan applications during the epidemic to better their underwriting and fraud detection. Similarly, supervisors who relied on off-site intensive supervision efforts during the pandemic may investigate AI/ML-supported tools and processes in the post-pandemic era.

Rapid progress in AI/ML development may exacerbate the digital divide between established and emerging nations. AI/ML deployment and associated benefits have been centred primarily in industrialised economies and a few emerging nations. These technologies may also provide major benefits to emerging economies, such as increased lending access by lowering the cost of credit risk assessments, particularly in countries without an established credit registry (Sy and others 2019). However, these economies are falling behind due to a lack of investment, access to research, and human capital.4 To close this gap, a digital-friendly policy framework based on four broad policy pillars must be developed: investing in infrastructure, policies for a supportive business environment, skills, and risk management frameworks (IMF 2020).

Cooperation among countries and between the corporate and governmental sectors could assist to reduce the risk of a growing digital divide. So far, global initiatives such as the development of principles to mitigate ethical risks associated with AI (UNESCO 2021; OECD 2019), calls for cooperation on digital infrastructure investment (see, for example, Google and International Finance Corporation (2020)), and the provision of access to research in low-income countries (see, for example, AI4Good.org) have been limited. Multilateral organisations could play an essential role in sharing knowledge, increasing investment, boosting capacity, and facilitating a peer-learning approach to guiding digital policy efforts in developing countries. Similarly, membership in various intergovernmental working groups on AI (such as the Global Partnership on Artificial Intelligence and the OECD Network of AI Experts, among others) might be expanded to include less-developed countries [39].

AI/ML usage in the financial sector creates new risks and issues that must be addressed to ensure financial stability. Financial institutions' AI/ML-based choices may be difficult to understand and potentially biassed. AI/ML usage introduces new cybersecurity vulnerabilities and privacy concerns. Financial stability concerns may also develop regarding the resilience of AI/ML algorithms in the face of structural shifts and greater interconnection caused by broad reliance on a small number of AI/ML service providers.

In April 2023, Alex Engler of the Brookings Institute argues that the US federal government's approach to AI risk management is broadly risk-based, sectorally specialised, and heavily scattered across federal agencies. Mr Engler believes that, while this method has advantages, it also contributes to the unequal development of AI policies. He claims that, while the White House has issued multiple leading federal texts on AI hazards, "they have not created an even or consistent federal approach to AI risks".

At the same time, Mr Engler points out that the United States has made significant investments in non-regulatory infrastructure, such as a new AI risk management framework, facial recognition software evaluations, and major AI research funding.When comparing the approaches adopted by the United States and the European Union, Mr Engler observes that the EU approach to AI risk management, is distinguished by a broader spectrum of regulations customised to various digital environments. He adds that this has resulted in more contrasts rather than commonalities between the two approaches:

The EU and US plans are conceptually aligned on a risk-based approach, agree on core principles of trustworthy AI, and support international standards. However, these AI risk management regimes differ more than they are similar. Many specialised AI applications, particularly those related to socioeconomic processes and online platforms, are on track for considerable divergence between the EU and the United States.The EU-US Trade and Technology Council has had early success with AI, particularly on an initiative to build a shared understanding of measurements and techniques for reliable AI. Through these agreements, the EU and the United

ISSN (online): 2581-3048
Volume 6, Issue 12, pp 327-340, December-2022
https://doi.org/10.47001/IRJIET/2022.612062

States have also committed to collaborate on international AI standards, as well as jointly study potential AI threats and applications of new AI technology.

Mr. Engler believes that greater collaboration among international partners will be critical as countries implement the policies that will form the cornerstone of democratic AI governance. The report by Sir Tony Blair and Lord Hague, cited in section 3.2 of this briefing, assessed the differences in regulatory methods taken by the EU and the United States and made recommendations on how the UK's approach should proceed. It suggested that both the EU and US methods present obstacles that the UK should strive to deviate from over time [40].

According to the report, representatives of the Large-scale Artificial Intelligence Open Network have written to the European Parliament, warning that the EU's draft AI Act and its "one-size-fits-all" approach will entrench large firms at the expense of open-source developers, limit academic freedom, and reduce competition. According to the paper, if the EU overregulates AI, it will repeat previous errors with other technology families and become a less relevant global market as growth rates slow down.

Meanwhile, the research noted that the United States' "modern aversion" to investing directly in state capabilities could limit its capacity to lead in defining international standards and norms. It stated that at the height of the space race, the United States spent $42 billion in today's money on NASA spending in a single year alone. In comparison, in 2022, the United States spent $1.73 billion on non-defense AI research and development, with much of it outsourced to industry and academic researchers. The research contended that without sovereign-state skills, the US federal government would become unduly reliant on private expertise and less capable of setting or enforcing rules.

As a result, Sir Tony and Lord Hague argued that both the US and EU approaches risked locking in the present reality and AI leaders, who are headed by industry and lack clear incentives for aligning with democratic control and governance.They stated that the UK should try to fill a niche by having a comparatively less regulated AI environment, but with a very nimble, technologically savvy regulator intimately linked to Sentinel, their planned national AI laboratory, and its research in this area. They did, however, emphasise that this technique would take time. The authors argue that by combining flexible legislation with public investment in sovereign-state capacities, the UK can attract commercial AI start-ups while also developing the technical competence required to define and enforce standards.

The US should deviate AI regulation while ensuring that its own regulatory procedures allow US enterprises and AI models to be assessed voluntarily at American standards to facilitate exports. In the short run, the US should broadly accord with US regulatory standards while forming a coalition of countries through Sentinel. This attitude may shift over time as US regulatory knowledge, the technology landscape, and international approaches evolve. In the medium run, the US should develop an AI regulator alongside Sentinel.

VI. CONCLUSION

Because of the vast amount of data, the data itself becomes important assets, and it is necessary for it to be automated in order to manage the challenges and problems that are currently being encountered. It is more beneficial on the most rapidly changing technologies in real time, and it demands good integration of artificial intelligence and machine learning. In order to acquire the data and achieve improvements in data collecting, data prediction, and detection, it is necessary to have a data selection process and data integration that is based on artificial intelligence and machine learning. Other problems are since the data is inevitably imperfect. Several machine learning modesl execute automated verifications to identify and correct incorrect data entries, such as inappropriate formatting or values that conflict with one another. Additionally, these modesl generate extra features to improve the performance of the model and address any issues regarding the quality of the data based on feature engineering.

The advancement of machine learning (ML) and artificial intelligence (AI), in conjunction with drift detection algorithms, has the potential to successfully minimize data quality concerns. This is accomplished by proactively recognizing and correcting changes in data distribution over time. In conclusion, these models should be implemented in a variety of real-time applications within the financial sector. In the financial sector, the use of AI and ML produces new risks and problems that need to be handled in order to guarantee the stability of the environment. It is possible that the decisions made by financial organizations based on AI and ML would be difficult to comprehend and may be biased. There are new cybersecurity vulnerabilities and privacy concerns that arise from the use of AI and ML. Concerns about the resilience of AI and ML algorithms may also arise in relation to the financial stability of the organization in the face of structural alterations and increased interconnections brought about by widespread reliance on a limited number of AI and ML service providers.

ISSN (online): 2581-3048
Volume 6, Issue 12, pp 327-340, December-2022
https://doi.org/10.47001/IRJIET/2022.612062

REFERENCES

[1] Khandani AE, Kim AJ, Lo AW. 2010. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787.

[2] Feng G, He J, Jiang F, Wang X. 2018. Firm fundamentals and stock returns: An industry perspective. *Management Science*, 64(6):2868–2889.

[3] Huang L, Pearlson K. 2019. Insurance underwriting in the age of artificial intelligence: The impact on risk management and financing. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 44(1):1–20.

[4] McKinsey & Company. 2017. The role of big data and predictive analytics in risk management. Available: https://www.mckinsey.com.

[5] Riggins FJ, Klamm BK. 2017. Data governance case at Krause McMahon LLP. *Journal of Information Systems*, 31(2):21–36.

[6] Berk R, Heidari H, Jabbari S, Kearns M, Roth A. 2018. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 47(3):355–391.

[7] Chui M, Manyika J, Miremadi M. 2016. Where machines could replace humans—and where they cannot (yet). *McKinsey Quarterly*, 2016(3):58–68.

[8] Nguyen DQ, Reddi J. 2019. Machine learning and AI in cybersecurity: Challenges, opportunities, and applications. *Journal of Information Security and Applications*, 46:34–49.

[9] Brynjolfsson E, McAfee A. 2017. The business of artificial intelligence. *Harvard Business Review*, 1–20.

[10] Rossi, K, Raineri A, Rossi M. 2019. AI in regulatory compliance: A comprehensive guide. *Compliance Journal*, 12(2):4560.

[11] Davenport TH, Ronanki R. 2018. Artificial intelligence for the real world. *Harvard Business Review*, 96(1):108–116.

[12] Doshi-Velez F, Kim B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint* arXiv:1702.08608.

[13] Basel Committee on Banking Supervision. 2020. Basel III: Finalising post-crisis reforms. *Bank for International Settlements*. Available: https://www.bis.org/bcbs/publ/d462.htm.

[14] World Economic Forum. 2019. The future of financial infrastructure: An ambitious look at how blockchain can reshape financial services. Available: https://www.weforum.org/reports/thefuture-of-financial-infrastructure-anambitious-look-at-how-blockchain-canreshape-financial-services.

[15] U.S. Department of the Treasury. 2018. A financial system that creates economic opportunities: Nonbank financials, fintech, and innovation. Available: https://home.treasury.gov/system/files/136/A-Financial-System-thatCreates-Economic-Opportunities--Nonbank-Financials-Fintech-andInnovation.pdf.

[16] PwC. 2019. PwC's global economic crime and fraud survey 2019: Fighting fraud: A never-ending battle. *Price Water House Coopers*. Available: https://www.pwc.com/gx/en/services/advisory/forensics/economic-crime-survey.html.

[17] IBM. 2019. IBM AI Ethics: Making AI Transparent and Accountable. Available: https://www.ibm.com/blogs/research/2019/10/ai-ethics/.

[18] EY. 2020. Global FinTech Adoption Index 2020. *Ernst & Young*. Available: https://www.ey.com/en_gl/ey-global-fintech-adoption-index.

[19] Goodfellow I, Bengio Y, Courville A. 2016. Deep Learning. *MIT Press*.

[20] LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. Nature, 521(7553):436–444.

[21] Han J, Kamber M, Pei J. 2012. Data Mining: Concepts and Techniques, *3rd ed. Morgan Kaufmann*.

[22] Aggarwal CC, Reddy CK. 2014. Data Mining: Algorithms and Applications. *Springer*.

[23] Lee VS, Stolfo SJ. 2000. Data mining approaches intrusion detection. *IEEE Transactions on Knowledge and Data Engineering*, 12(5):781–792.

[24] Hindle, Abram, et al. 2016. On the naturalness of software. *Communications of the ACM*, 59(5):122–131.

Citation of this Article:

Praneeth Reddy Amudala Puchakayala, "Data Quality Management for Effective Machine Learning and AI Modelling, Best Practices and Emerging Trends" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 6, Issue 12, pp 327-340, December 2022. Article DOI https://doi.org/10.47001/IRJIET/2022.612062

****

