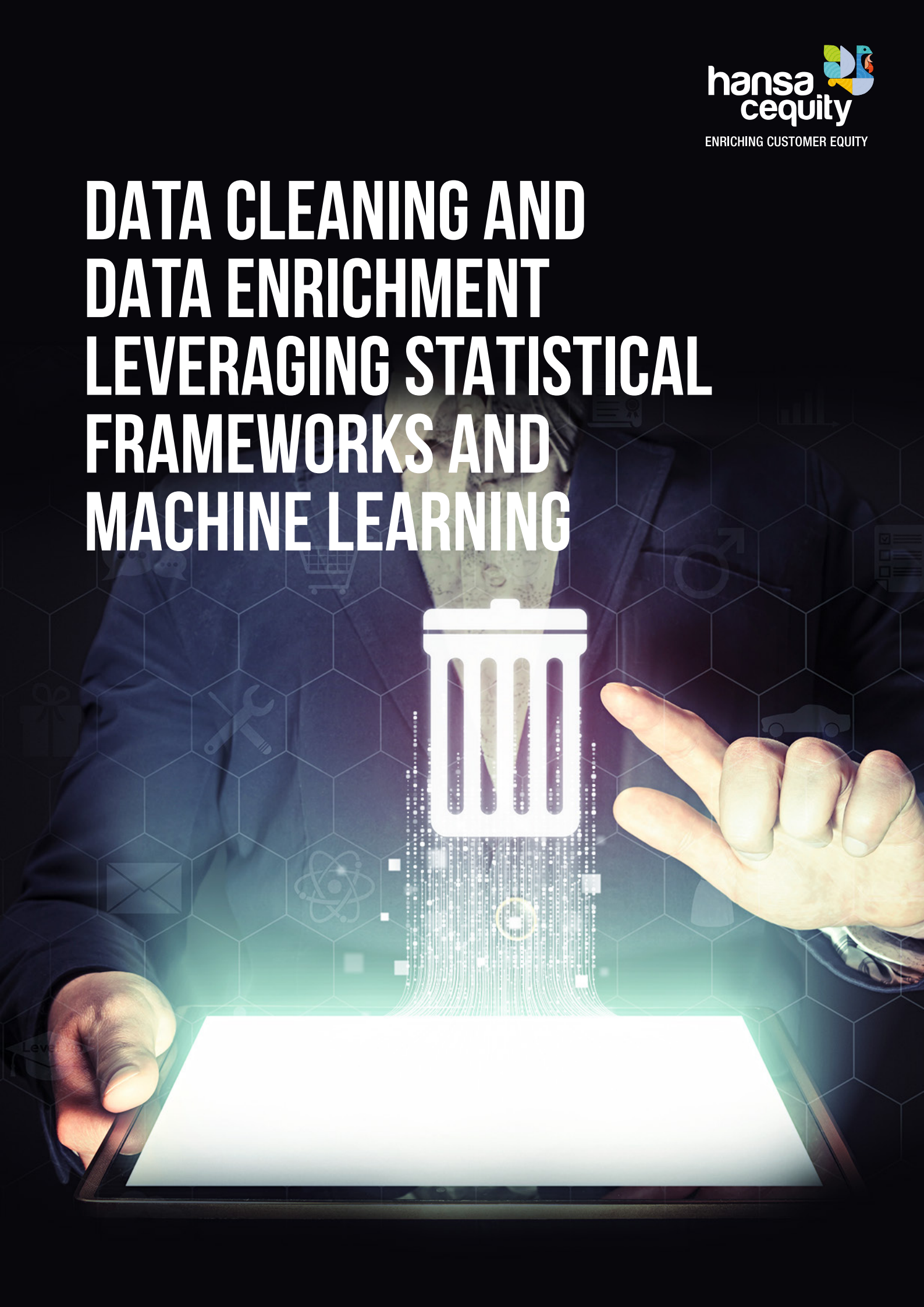
hansa cequity

ENRICHING CUSTOMER EQUITY

DATA CLEANING AND DATA ENRICHMENT LEVERAGING STATISTICAL FRAMEWORKS AND MACHINE LEARNING



Data Cleaning and Data Enrichment leveraging Statistical Frameworks and Machine Learning

5 REASONS WHY DATA CLEANING IS IMPORTANT:

1. Improve ROI of marketing campaigns
2. Targeting the right customers with relevant messaging
3. Reduce overall costs
4. Remove duplicates and improve CRM efficiency
5. Make better business decisions

SYNOPSIS:

In today's data-driven world, the quality of data plays a pivotal role in the success of any organization. Reliable and accurate data is the foundation upon which effective decision-making, business intelligence, and advanced analytics are built. However, the reality is that raw data is often messy, incomplete, and error-prone. This is where data cleaning, a crucial component of the data pre-processing phase, comes into play. In this article, we will explore the best data cleaning practices in the industry, focusing on the integration of data science and statistics to ensure high-quality datasets. We give an overview of all the data cleaning & data enrichment methods leveraging open-source tools, statistical techniques, and new-age deep learning methods such as GANs (Generative Adversarial Networks).

AREAS OF FOCUS:

I. Understanding the Importance of Data Cleaning

II. Data Cleaning Frameworks

III. Dealing with Missing Data

IV. Outlier Detection and Treatment

V. Addressing Duplicates

VI. Validation through Descriptive Statistics

VII. Anomaly Detection

VIII. GANs (Generative Adversarial Networks) and Deep Learning for Data Enrichment

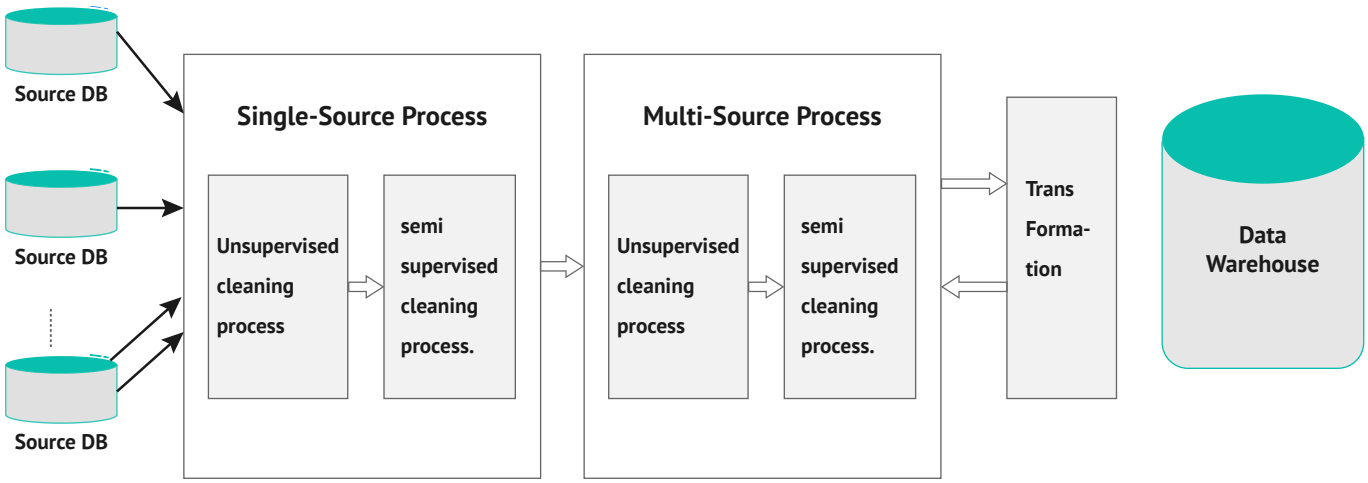
UNDERSTANDING THE IMPORTANCE OF DATA CLEANING:

Before delving into specific practices, it's essential to grasp the significance of data cleaning. Dirty data can lead to flawed analyses, inaccurate insights, and ultimately poor business decisions through Decision Support Systems which executives/CXOs rely upon. By cleaning, enriching, and preparing data effectively, organisations can enhance the reliability of their findings, reduce errors, and save valuable time and resources for better targeting, accurate marketing while building deep customer engagement, brand trust, and loyalty. The clean and well-enriched data can also be a key driver for higher CLTV (Customer Lifetime Value) and brand equity.

DATA CLEANING FRAMEWORKS:

Data cleaning is not a one-size-fits-all process. Different datasets require tailored approaches. Several frameworks guide data scientists and analysts through the data-cleaning journey. The widely used frameworks include Tidy Data Principles:

Based on the work Hansa Cequity has done for more than a decade, the Tidy Data Principles emphasise organising data in a structured, standardised format called CSV - Customer Single View. This simplifies the cleaning process, making it easier to identify and rectify issues. Our Python experts leverage Pandas, a powerful data manipulation library in Python, that provides functions and methods for efficient data cleaning. It allows users to handle missing values, duplicate entries, and outliers seamlessly. Our R experts, the dplyr and tidyr packages offer similar functionality to Pandas. They provide a grammar of data manipulation, enabling users to filter, arrange, and clean data easily.



DEALING WITH MISSING DATA:

One of the most common challenges in data cleaning is handling missing values. Several statistical and data science techniques can be employed:

Imputation Methods:

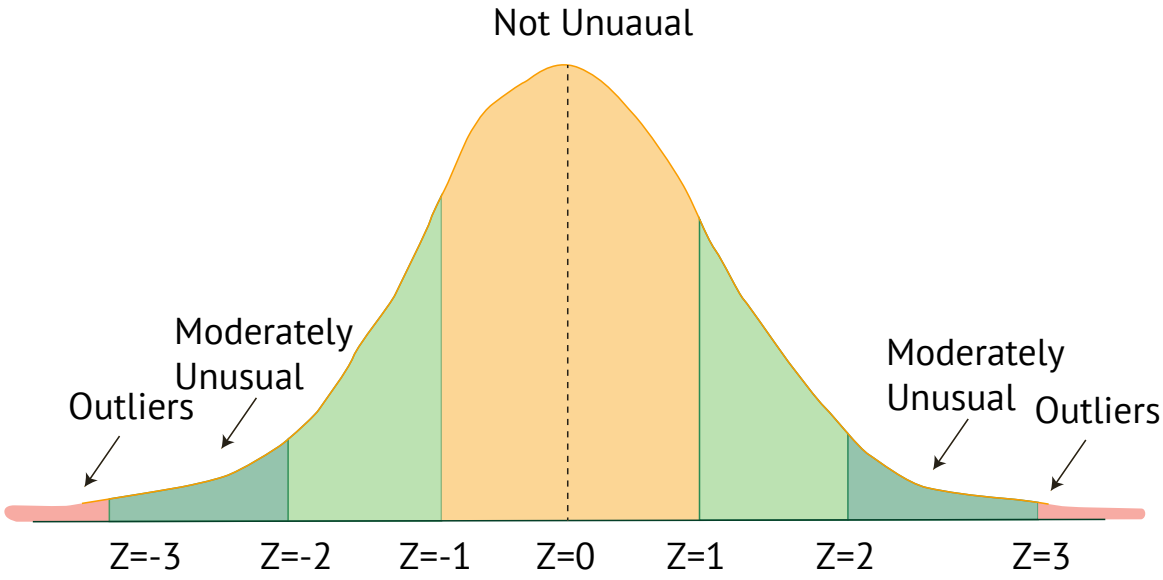
Statistical imputation methods, such as mean imputation, regression imputation, and K-Nearest Neighbors (k-NN) imputation, are employed to estimate and fill in missing values based on the characteristics of the existing data.

Understanding the Missing Data Mechanism:

Identifying the pattern of missing data (missing completely at random, missing at random, or missing not at random) helps in selecting the appropriate imputation strategy. Our experts deploy important statistical techniques such as multinomial regression, and machine learning methods such as XGBoost and Neural Networks to impure the missing data as per the pattern of missingness unless the missing data is completely random.

OUTLIER DETECTION AND TREATMENT:

Outliers can significantly skew analysis results and impact the business decisions driven by the Decision Support Systems which are dependent on the data. Leveraging statistical techniques and machine learning algorithms our experts effectively identify and manage outliers:



Z-Score and IQR Methods:

Z-Score and Interquartile Range (IQR) methods help our experts to identify and remove outliers by considering the standard deviation or distribution of the data along with our domain knowledge.

Machine Learning-Based Outlier Detection:

Our data science experts leverage machine learning algorithms like Isolation Forests, One-Class SVM, and Local Outlier Factor (LOF) can be applied to more complex datasets where traditional methods may fall short. These algorithms function extremely well on any missing data increasing the accuracy, precision, and recall of the models in a given industry.

ADDRESSING DUPLICATES:

Duplicate entries in a dataset can compromise its integrity. While we leverage first-party data including demographic information and third-party data based on our custom crawlers, our experts also employ both statistical and algorithmic approaches that can help identify and eliminate duplicates:

Exact and Fuzzy Matching:

Exact matching using unique identifiers and fuzzy matching algorithms like Levenshtein distance can identify and handle duplicates with slight variations. We also leverage some of the algorithms taken from other domains such as approximate string matching and genetic matching for fuzzy matching. This makes sure that the data we work upon is deduped to the optimum level reducing the marking costs.

Data Profiling:

Our data experts have mastered the art of comprehensive data profiling which involves analyzing data distributions, patterns, and uniqueness to detect potential duplicates. Data Profiling is the process of examining, analyzing, and creating summarized version through various statistical modeling approaches which showcase any issues and risks in the data with respect to industry benchmarks and quantitative analysis of the structured and unstructured data. This gives us an edge to deal with duplicate entries in any given customer data.

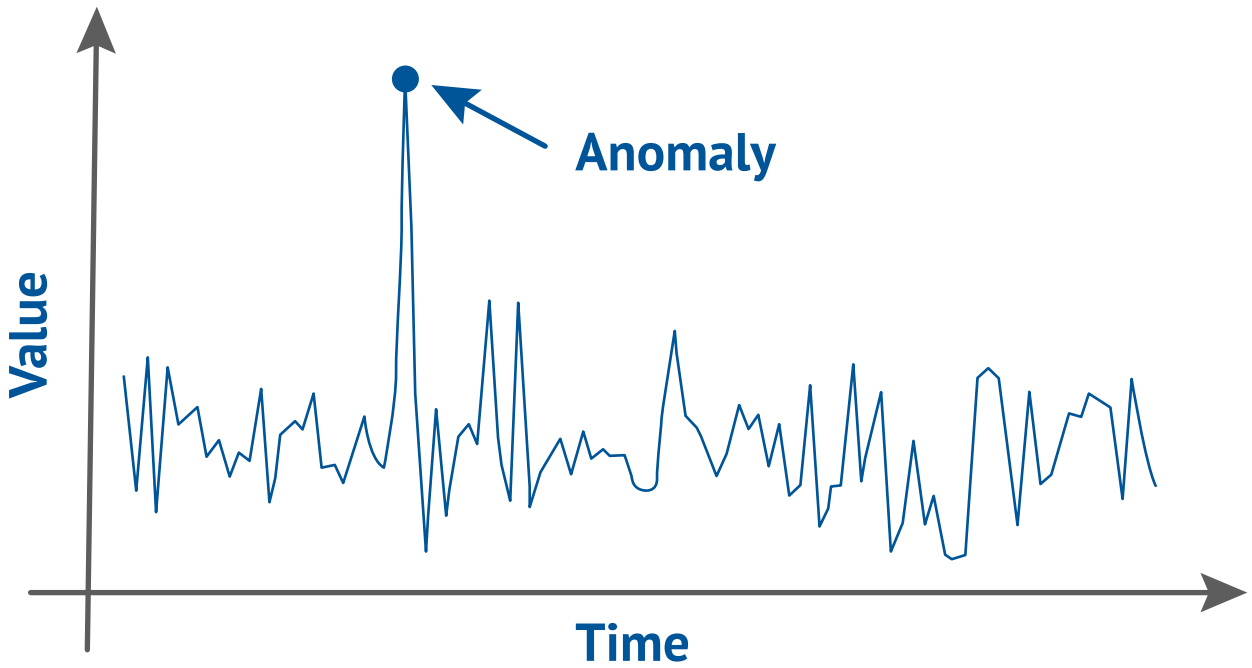
VALIDATION THROUGH DESCRIPTIVE STATISTICS:

Utilizing descriptive statistics is crucial for validating the effectiveness of data cleaning processes. Calculating summary statistics such as mean, median, standard deviation, and skewness before and after cleaning can reveal the impact on the dataset's distribution. We also understand different processes and what distribution the data in a particular industry might follow. E.g. Pharma drug discovery process will have skewed right-tail distribution with targets on the right tail of the distribution vs retail shopping data will be mostly in normalized distribution with spikes that are yielded by different successful or unsuccessful campaigns.

ANOMALY DETECTION:

Anomaly detection, a crucial facet of data science, involves the identification of irregular patterns or outliers within datasets, and it finds applications across various domains, including finance, cybersecurity, and industrial monitoring.

Data scientists at Hansa Cequity employ an array of techniques, prominently utilizing machine learning algorithms to sift through vast datasets and pinpoint instances that deviate significantly from the norm. Unsupervised learning methods, such as clustering algorithms (e.g., k-means) and density-based techniques (e.g., DBSCAN), excel in detecting anomalies without the need for labeled data, making them versatile tools for various industries. These algorithms assess data points based on their relationships, densities, or distances from the majority, effectively isolating anomalies and raising alerts when unexpected patterns emerge.



In our experience, Support Vector Machines (SVMs) stand out as effective tools in anomaly detection by defining optimal hyperplanes that separate normal data points from potential outliers. SVMs excel in high-dimensional spaces, making them suitable for complex datasets. Isolation Forests, another powerful algorithm, takes a different approach by constructing random decision trees and isolating anomalies based on their shorter average path lengths. This method is particularly efficient for detecting anomalies in large datasets and is resilient to the curse of dimensionality. By leveraging the strengths of SVMs and Isolation Forests, our data scientists are able to enhance the anomaly detection capabilities, ensuring accurate identification of irregularities amidst vast and complex datasets.

In recent years, deep learning techniques have revolutionized anomaly detection, particularly with the advent of autoencoders. Autoencoders are neural network architectures designed to learn compact representations of input data. During the training phase, these models encode normal patterns, and any deviation from this learned representation is flagged as an anomaly during the testing phase. This approach is particularly effective in capturing intricate, non-linear relationships within data, making it suitable for applications where anomalies may manifest as subtle deviations from the norm. By integrating deep learning techniques like autoencoders into anomaly detection workflows, data scientists can elevate their capacity to uncover complex patterns and enhance the robustness of anomaly detection systems across diverse industries.

GANS (GENERATIVE ADVERSARIAL NETWORKS) AND DEEP LEARNING FOR DATA ENRICHMENT:

Data enrichment is a critical process in enhancing the quality and depth of datasets, and Generative Adversarial Networks (GANs) have emerged as a powerful tool in this domain. GANs, a class of machine-learning models, are renowned for their ability to generate synthetic data that closely resembles real-world samples. In the context of data enrichment, GANs can be employed to augment existing datasets by generating additional instances that capture the underlying patterns and structures present in the original data. By

training a GAN on a given dataset, it learns to create realistic data points that mimic the distribution of the input data. This generated data can then be seamlessly integrated with the original dataset, addressing issues of data scarcity and diversity. The use of GANs in data enrichment not only expands the volume of available data but also contributes to improved model generalization, robustness, and performance across various applications in machine-learning and data analytics. Our data scientists have figured out a way to produce GANs with optimum convergence which delivers a perfect enriched data outcome.

CONCLUSION

In conclusion, effective data cleaning is indispensable for extracting meaningful insights from raw datasets. By integrating data science and statistical techniques, organizations can ensure the reliability and accuracy of their data, laying a solid foundation for informed decision-making.

As technology advances, the importance of these practices will only grow, emphasising the need for continuous improvement and adaptation in the dynamic landscape of data analytics. Embracing these best practices will not only enhance data quality but also empower organisations to unlock the full potential of their data assets.

CONTRIBUTORS

Prasad Kothari
Head -
Data science & AI

Trideep Goswami
Associate Director -
Marketing Solutions

Amit Panjani
Senior Consultant -
Data Solutions Group

To know more, contact us at marketing@hansacequity.com

hansa cequity
ENRICHING CUSTOMER EQUITY

Consulting I Data Management
Analytics & Insights I Campaign Management
Digital Experience I Customer Experience

403 & 404, B Wing,
4th Floor, Commercial Office Towers,
Kohinoor City Mall, Kirol Road, Off LBS Marg,
Kurla (W), Mumbai 400 070.
Email: info@cequitysolutions.com
http://www.hansacequity.com

Mumbai, Delhi, Bangalore, Chennai and Chicago.