Big Data Mining Process and Application

Table of Contents

[A 3](#_Toc136704068)

[Introduction 3](#_Toc136704069)

[Overview 3](#_Toc136704070)

[Conclusion 5](#_Toc136704071)

[References 5](#_Toc136704072)

[B 6](#_Toc136704073)

[Introduction 6](#_Toc136704074)

[Business Understanding 6](#_Toc136704075)

[Data Understanding 6](#_Toc136704076)

[Data Preparation 7](#_Toc136704077)

[Modeling Techniques for Fraud Detection 7](#_Toc136704078)

[Evaluation and Performance Metrics 8](#_Toc136704079)

[Deployment and Implementation Success Criteria 8](#_Toc136704080)

[Measurable Success Criteria 9](#_Toc136704081)

[Conclusion 9](#_Toc136704082)

[References 9](#_Toc136704083)

# A

Title: A Critique of "The CRISP-DM Model: The New Blueprint for Data Mining" in the context of Big Data Mining in 2023

## Introduction

Business understanding, data understanding, data preparation, modelling, evaluation, as well as deployment are the six phases that make up the Cross-Industry Standard Process for Data Mining (CRISP-DM), as outlined in "The CRISP-DM Model: The New Blueprint for Data Mining" by Shearer (2000). The paper by Shearer is aptly named "The New Blueprint for Data Mining." This study aims to synthesise two recent journal articles (published in 2018 or later) that examine the CRISP-DM method of data mining and determine if it is still relevant and useful in contemporary 'Big Data' mining scenarios. The authors of both pieces are academics with an interest in the subject.

## Overview

When it was first proposed, the CRISP-DM model was hailed as a revolutionary breakthrough due to the fact that its methodology is rigorous, systematic, and can be generalised. Data miners had access to a comprehensive resource that assisted them in effectively addressing data mining issues in a variety of different contexts thanks to the existence of this resource. Shearer (2000) explains how the sequential yet iterative method satisfied the needs of both business and technical stakeholders by striking a balance between flexibility and structure. This allowed the method to meet the needs of all of the stakeholders involved. Because of this, everyone was on the same page, they were aware of their responsibilities, and they were able to successfully complete the project by working together.

On the other hand, when viewed in the context of the data landscape of today, particularly that of "Big Data" in the year 2023, significant flaws in the CRISP-DM strategy have become readily apparent. To begin, the 'Big Data' phenomenon of exploding data volume, velocity, and variety makes the conventional data preparation stage of the CRISP-DM approach inherently more difficult to complete. This is where the approach begins. According to Vijay Kotu and Bala Deshpande (2018), this has resulted in the development of new methods, such as automated data cleaning and preparation pipelines. These methods were made possible as a direct result of the aforementioned.

This argument is supported by a recent study that was carried out by Bhardwaj and Pal (2020), who postulate that CRISP-DM struggles to cope with the sheer magnitude of Big Data. This argument is supported by a recent study that was carried out by Bhardwaj and Pal (2020). The findings of a recent research project that was carried out by Bhardwaj and Pal (2020) lend credence to this argument. They shed light on the fact that in order to successfully manage the data preparation stage of large-scale data mining projects, more dynamic, scalable, and automated approaches are required.

Second, the use of complex machine learning and deep learning models for mining 'Big Data' calls for a more nuanced understanding and approach to the modelling stage of the process. For example, the CRISP-DM model does not explicitly address crucial aspects of improving model performance such as parameter optimisation and model tuning (Kelleher, Mac Namee, & D'Arcy, 2015). These aspects are both vital to improving model performance.

Stankovic, Suknovic, and Milosavljevic (2018) write convincingly about the deficiencies of CRISP-DM in the context of the era of Big Data in their research paper. They propose extending the capabilities of the CRISP-DM methodology in order for it to be able to take into account cutting-edge methodologies such as deep learning and reinforcement learning. These considerations highlight that even though the core of the CRISP-DM model is still relevant, adaptations are necessary in order to keep pace with evolving data analytics paradigms in order to maintain the model's relevance. This is the case even though the core of the model is still relevant.

Despite these limitations, the CRISP-DM model's strength is that it can be used in any problem-solving circumstance and follows a systematic approach. It lays a firm groundwork for data mining, and with some tweaks to the existing paradigm, it can remain a useful framework in the age of big data as well.

Shearer's (2000) "The CRISP-DM Model: The New Blueprint for Data Mining" essay introduced a fresh, systematic, and standardised strategy for data mining. The acronym "CRISP-DM" stands for the Commonly Accepted Reference Information Set Process for Data Mining. The sum of its six stages—"business understanding," "data understanding," "data preparation," "modelling," as well as "evaluation and deployment"—are iteratively progressed through. However, it is vital to evaluate whether or not the CRISP-DM method will still be relevant in 2023 in light of the big data revolution as well as the substantial changes it has caused to the world of data analytics. In this analysis, we will look at the original publication in addition to two other articles published in 2018 or later that share similar themes.

Shearer as well as his approach have contributed significantly to the standardisation of the data mining process. The framework of the model is both adaptable and iterative, and as a result, it has seen widespread adoption across a variety of industries. In addition to this, it has made effective communication between technical and non-technical stakeholders much easier to achieve, which has resulted in the alignment of the data mining technical process with business goals.

Despite this, the advent of big data, which is distinguished by massive volume, velocity, and variety, has forced the CRISP-DM model to contend with a plethora of challenges, particularly in the phase of data preparation. According to Vijay Kotu and Bala Deshpande (2018), traditional methods of data cleaning and preprocessing may be adequate for processing smaller datasets; however, big data processing calls for methods that are more automated and scalable.

Bhardwaj and Pal (2020) agree, saying that the CRISP-DM approach has been useful in the past but now has trouble keeping up with the volume and variety of information found in big data. The fact that the CRISP-DM method has been successful in the past lends credence to this claim. The authors stress that CRISP-DM's data preparation step, which is manual and somewhat labour expensive, does not match the speed as well as automation needs of big data analytics. This is because CRISP-DM requires human intervention at the data preparation stage.

The emergence of sophisticated machine learning and deep learning algorithms adds a new layer of difficulty to the modelling process. Parameter adjustment and model selection for these algorithms are sometimes complex and not covered in the basic CRISP-DM model. In many cases, these methods are essential for excelling. This void is highlighted by Kelleher, Mac Namee, and D'Arcy (2015), who stress the importance of algorithm-specific treatment in the modelling stage.

Similar to what has been said above, Stankovic, Suknovic, and Milosavljevic (2018) point out the shortcomings of the CRISP-DM model in the era of sophisticated machine learning algorithms and massive amounts of data. They present a refined version of the CRISP-DM procedure that incorporates reinforcement learning and other modern machine learning as well as deep learning techniques. Their findings show that CRISP-DM has to be revised to accommodate the new technologies and philosophies that are rapidly altering the data analytics landscape.

Despite these challenges, it is crucial to remember that the CRISP-DM model's fundamental strength rests in the widespread applicability of its methodology, the systematic nature of its approach, as well as the structured manner in which it solves problems. Despite certain limitations, it is still a vital tool in the data mining industry. This model needs to be updated and revised so that it can accommodate the demands of big data and sophisticated analytics.

## Conclusion

It may be concluded that the CRISP-DM model has made a substantial contribution to data mining; however, the dynamics of huge data need the creation of a more sophisticated model. To meet the problems of big data and fully realise its potential, a technique that is more scalable, automated, as well as intelligent is required in this era of big data. In order to efficiently integrate more sophisticated tools and methods, automate preprocessing stages, and build in dedicated facilities for model tweaking and optimisation, further study should focus on refining the CRISP-DM procedure.

The CRISP-DM model has proven to be a powerful and effective tool for data mining. However, the model requires revision so that it can incorporate new information and techniques, such as "Big Data" and complex algorithms for machine learning. Future studies need to concentrate on improving the CRISP-DM framework by integrating more complex methodologies, automated preprocessing procedures, as well as explicit steps for model tuning as well as optimisation.

References:

1. Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, 5(4), 13-22.

2. Vijay Kotu & Bala Deshpande (2018). Data Science Concepts and Practice. Elsevier.

3. Bhardwaj, B. & Pal, S. (2020). Big Data: An exploration of its relevance to the CRISP-DM methodology. Big Data & Society, 7(1), 1-10.

4. Kelleher, J.D., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of Machine Learning for Predictive Analytics. MIT Press.

5. Stankovic, M., Suknovic, M., & Milosavljevic, M. (2018). Towards Enhanced Predictive Analytics: An Extended CRISP-DM Process for Big Data Domain. Computer Science and Information Systems, 15(2), 255-275.

# B

Title: Big Data Mining Case Study: Fraud Detection in Financial Transactions

## Introduction

Businesses as well as financial institutions have a huge difficulty when dealing with fraudulent activity in financial transactions. Minimising financial losses, maintaining consumer confidence, and meeting regulatory obligations all hinge on being able to detect and prevent fraud. Because of the volume and variety of transaction data, it is essential to employ big data mining strategies in order to identify fraud efficiently.

The financial and reputational costs of fraud to firms may be enormous. Avoiding losses and protecting the company's reputation depends on the prompt identification of fraudulent transactions.

In order to detect trends, outliers, and other signs of fraudulent behaviour, firms might use big data mining techniques to examine massive amounts of transactional data. Businesses may improve their capacity to detect and prevent fraud by using sophisticated analytics and machine learning algorithms (Hasan et al., 2020).

## Business Understanding

Spotting and stopping fraudulent financial transactions is a top priority for every firm. The goal is to reduce financial losses, safeguard customers' interests, and preserve the company's good name by spotting any instances of fraud, theft, or other questionable activity.

Due to the dynamic nature of fraudulent actions, detecting them might be difficult. Because fraudsters frequently modify their methods, conventional rule-based systems can only do so much. Manually spotting fraudulent trends is further made difficult by the sheer volume as well as complexity of transaction data.

The obstacles of fraud detection can be surmounted with the use of big data mining techniques. Businesses may analyse massive datasets using machine learning algorithms, anomaly detection, as well as pattern identification to find fraud trends and questionable actions that would otherwise go undetected .

## Data Understanding

The transactional data utilised for fraud detection includes a wide range of parameters, such as the amount, kind, account information, timestamps, and more. This information may be used to learn more about lawful transactions and identify any signs of fraudulent activity.

Companies may improve their fraud detection by utilising other data sources in addition to transactional data. Blacklists of known fraudsters, historical fraud data, industry-wide fraud indicators, as well as other pertinent data can give extra context and insights for recognising suspicious actions (Zhou et al., 2020).

Understanding the transactional data relies heavily on data exploration methods like descriptive statistics, data visualisation, and exploratory data analysis. These methods aid in determining where problems exist in terms of data quality, distribution patterns, outliers, and possible correlations between variables, all of which guide further data preparation and modelling processes.

## Data Preparation

Cleaning up transactional data include locating and dealing with missing values, eliminating duplication, and fixing inconsistencies and mistakes. To ensure the data is in an appropriate format for analysis, preprocessing operations may include involve normalising, scaling, as well as encoding categorical variables.

The generation of new features or changes of existing features to increase the predictive capacity of the models is called "feature engineering," and it is an essential part of the fraud detection process. Statistics, data aggregation, time-based features, and the addition of new data sources are all possibilities.

Transformation and normalisation methods are used to make the variables consistent and comparable. Depending on the nature of the variables and the needs of the selected modelling methodologies, appropriate transformations may involve logarithmic transformations, standardisation, or min-max scaling .

## Modeling Techniques for Fraud Detection

Using historical data and labelled samples, supervised learning algorithms are frequently employed in fraud detection to categorise transactions as fraudulent or lawful (Eltweri et al., 2021). The following are examples of well-known classification systems used in detecting fraud:

Decision trees build a structured model using a set of rules for making decisions at each level. They are readable and capable of capturing intricate interrelationships between characteristics. Based on the values of such attributes, decision trees may determine if a transaction is fraudulent or not.

The goal of Random Forest, an ensemble learning approach, is to increase precision and decrease overfitting by combining numerous decision trees. The final forecast is based on the majority vote of the individual decision trees, which are each trained on a random subset of the data.

In a high-dimensional feature space, Support Vector Machines (SVMs) create a hyperplane that divides legal from fraudulent transactions. Finding the appropriate decision border between the two classes is the goal of support vector machines (SVM) (Aboud et al., 2022).

Anomalies or outliers in the data that may suggest fraudulent activity can be discovered with the use of unsupervised learning methods. Anomaly detection algorithms can find new fraud trends even without labelled instances. The following are examples of two widely-used anomaly detection algorithms in fraud detection:

An anomaly-isolating technique, Isolation Forest uses a tree-based structure. The method works by first picking a feature at random, and then picking a split value at random from the feature's interval (Tang & Karim, 2019). Instances are kept segregated from the rest of the data by the algorithm until the anomalies have been found and removed.

An example of a density-based clustering technique, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clusters instances together according to their closeness and density. Outliers, or instances that do not fit into any cluster or are located in low-density zones, are discovered; these may be signs of fraudulent activity.

## Evaluation and Performance Metrics

Accuracy, precision, recall, F1-score, as well as area under the receiver operating characteristic curve (AUC) are some of the evaluation measures used to analyse the effectiveness of fraud detection algorithms. The accuracy metric assesses how well the model predicts the future, while the precision metric zeroes in on how well it identifies instances of fraud. The percentage of real fraudulent transactions that were accurately recognised by the model is the recall (or sensitivity). The F1-score is a balanced indicator of a model's efficacy, being the harmonic mean of accuracy and recall.

Estimating the effectiveness of fraud detection models on unseen data is a popular use of cross-validation methods like k-fold cross-validation. Cross-validation gives a reliable evaluation of the model's generalisation capacity by splitting the data into training and testing subsets as well as repeating the procedure numerous times with various subsets (Al-Hashedi & Magalingam, 2021).

Reducing the number of false positives (transactions that are genuine but are incorrectly labelled as fraudulent) as well as false negatives (fraudulent transactions that are incorrectly labelled as normal) is essential in fraud detection. Customers may become dissatisfied and needless research expenses may be incurred due to a false positive, while losses may occur due to fraud that goes unnoticed due to a false negative. In order to create reliable fraud detection models, it is crucial to strike a balance between these two types of mistakes.

## Deployment and Implementation Success Criteria

Integration with the preexisting transaction processing system is essential for the successful implementation of fraud detection algorithms. This connection paves the way for both real-time and batch processing of transactions, with the models analysing transaction data to raise red flags and trigger alarms.

With real-time deployment, any fraudulent activity during transaction processing may be stopped instantly. In contrast, executing models on a significant amount of historical transaction data in batches is what batch processing is all about. The needs and limitations of the organisation should guide the decision between real-time and batch processing.

Businesses may reap several benefits from implementing a big data mining-based fraud detection system (Herland et al., 2019). Some of these benefits include fewer monetary losses from fraud, happier customers, more regulatory compliance, and the ability to proactively spot new fraud trends. Maintaining the company's good name and safeguarding its stakeholders are additional benefits of investing in a solid fraud detection system.

## Measurable Success Criteria

Metrics like detection accuracy, precision, recall, F1-score, as well as AUC may be used to evaluate the effectiveness of a given fraud detection solution. The system's capacity to distinguish between fraudulent and legal transactions is indicative of higher accuracy and performance measures.

One measure of success is a reduction in fraud-related financial losses. By improving their methods of fraud detection and prevention, firms may reduce the negative financial impact of fraudulent transactions and save money.

The speed with which the system can react to identify and prevent fraud is another indicator of the success of the deployment. The effects of fraud can be mitigated and additional losses avoided by its rapid discovery and aggressive avoidance.

It is crucial for organisations to disclose any suspected instances of fraud as required by law. The efficacy of the solution is shown by the degree to which the laws are followed and fraudulent acts are reported promptly (Hasan et al., 2020).

## Conclusion

Increased precision, less monetary losses, increased consumer trust, and regulatory compliance are just some of the many positive outcomes that may result from incorporating big data mining techniques into fraud detection in financial transactions. To efficiently identify and prevent fraudulent activity, a complete methodology is needed, and this may be provided by the combination of supervised learning and unsupervised learning algorithms.

Detecting fraud is an ongoing process that requires the models to be constantly evaluated and modified to account for new types of fraud. Improve your organization's fraud detection skills and get ahead of developing fraudulent behaviours by reviewing performance indicators and comments from fraud investigations on a regular basis (Zhou et al., 2020).

# References

Hasan, M.M., Popp, J. and Oláh, J., 2020. Current landscape and influence of big data on finance. Journal of Big Data, 7(1), pp.1-17.

Zhou, H., Sun, G., Fu, S., Fan, X., Jiang, W., Hu, S. and Li, L., 2020. A distributed approach of big data mining for financial fraud detection in a supply chain. Comput Mater Continua, 64(2), pp.1091-1105.

Eltweri, A., Faccia, A. and KHASSAWNEH, O., 2021, December. Applications of Big Data within Finance: Fraud Detection and Risk Management within the Real Estate Industry. In 2021 3rd International Conference on E-Business and E-commerce Engineering (pp. 67-73).

Tang, J. and Karim, K.E., 2019. Financial fraud detection and big data analytics–implications on auditors’ use of fraud brainstorming session. Managerial Auditing Journal, 34(3), pp.324-337.

Al-Hashedi, K.G. and Magalingam, P., 2021. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. Computer Science Review, 40, p.100402.

Herland, M., Bauder, R.A. and Khoshgoftaar, T.M., 2019. The effects of class rarity on the evaluation of supervised healthcare fraud detection models. Journal of Big Data, 6, pp.1-33.

Aboud, A. and Robinson, B., 2022. Fraudulent financial reporting and data analytics: an explanatory study from Ireland. Accounting Research Journal, 35(1), pp.21-36.