IMPLEMENTATION OF DATA MINING TECHNIQUES IN THE PRIVATE SECTOR OF THE REPUBLIC OF MOLDOVA: CASE STUDIES AND COMPETITIVE ADVANTAGES

CZU: [004.6:005]:334.722(478)
DOI: https://doi.org/10.53486/tids2026.22

VITALIE URSACHI
PhD, University Lecturer, Academy of Economic Studies of Moldova
e-mail: ursachi.vitalie@ase.md
ORCID ID: 0009-0008-9206-1324

EMILIA GOGU
PhD, Associate Professor, Bucharest University of Economic Studies, Romania
e-mail: emilia.gogu@csie.ase.ro
ORCID ID: 0000-0002-7333-7128

Abstract. This paper examines the potential of Data Mining techniques to generate measurable competitive advantages in the private sector of the Republic of Moldova, drawing on three applied case studies conducted by the authors. The main methodological frameworks — CRISP-DM, KDD, SEMMA, TDSP, Scrum and Kanban — are compared, with an emphasis on the prevalence of CRISP-DM. The case studies include: a scoring model for debt collection (−27% ineffective actions), market value estimation for automobiles (R^2 = 0.96) and apartments (R^2 = 0.90) at the VIA SCOPE Bureau, and an energy loss forecasting system at RED-Nord (R^2 = 0.98; MAPE = 2.23%). The findings show that Data Mining projects structured according to CRISP-DM can generate measurable business value, while the case studies suggest that organisational and analytical capabilities may represent important barriers to wider adoption.

Keywords: Data Mining, CRISP-DM, machine learning, case studies, competitive advantages, Republic of Moldova.

JEL Classification: C38, C53, M15

1. INTRODUCTION

Data Mining is the process of discovering valuable models, patterns and knowledge in large datasets by combining methods from statistics, machine learning and database management. The term is closely linked to the KDD (*Knowledge Discovery in Databases*) process, introduced in 1989 at IJCAI, of which it constitutes the central analysis stage (Fayyad *et al.*, 1996). The basic stages are defined quite clearly by Bramer (Bramer, 2016). In the process described, Data Mining is only one stage of the overall process. It should also be noted that Data Mining is an interdisciplinary field: statistics, databases, pattern recognition, artificial intelligence and visualisation all contribute equally to knowledge extraction.

Artificial intelligence (AI), data science, data mining, machine learning and deep learning are interconnected yet distinct concepts. **AI** aims to create intelligent machines; **data science** extracts valuable insights; **data mining** uncovers hidden patterns; **machine learning** enables computers to learn from data; **deep learning** simulates neural networks (Rohan, 2023). Although widely adopted internationally, Data Mining remains underexploited in the Republic of Moldova. Through three case studies, this paper examines how Data Mining implementation can generate measurable competitive advantages and explores organisational and analytical factors that may constrain wider adoption.

The aim of the research is to assess how Data Mining projects structured according to the CRISP-DM methodology can generate measurable business value in the private sector of the Republic of Moldova.

2. THEORETICAL FRAMEWORK: DATA SCIENCE PROJECT MANAGEMENT METHODOLOGIES

Executing a Data Mining project requires a structured methodological framework that guides the team from problem definition through to solution deployment. A review of the literature identifies six main methodologies, presented in Table 1 (Mariscal *et al.*, 2010).

Table 1. Comparative analysis of data science project management methodologies

Methodology	Origin	Structure	Main characteristics
CRISP-DM	1996	6 iterative phases	Flexible, tool-independent; the most widely used in industry and research (~43% of practitioners)
SEMMA	2000	5 phases (SAS)	Oriented towards the SAS platform; limited portability outside the SAS ecosystem
KDD	1996	9 stages	Rigorous academic approach; conceptual precursor of modern data science
TDSP	2016	5 stages	Integrated with the Microsoft Azure ecosystem; suited to corporate teams
Scrum	2001	Iterative sprints	Maximum agility; requires role adaptation for DS projects
Kanban	2007	Continuous flow	Clear visualisation of progress; less structured for the modelling phase

Source: *compiled by the authors based on the specialised literature*

CRISP-DM (*CRoss Industry Standard Process for Data Mining*) is the dominant methodology worldwide: ~43% of practitioners use CRISP-DM as their main framework (Piatetsky, 2014, Schröer *et al.*, 2021). Moreover, a recent study shows that 86% of team members apply CRISP-DM frequently or always, starting from the observation that 82% of teams have no process model or methodology at all (Shimaoka *et al.*, 2025). Developed in 1996 by Daimler-Benz, OHRA, NCR and SPSS, it proposes a cyclical architecture with six interdependent phases (Table 2). Its iterative nature and tool independence are its key structural advantages (Chapman *et al.*, 2000, Martínez-Plumed *et al.*, 2021).

Table 2. Phases of the CRISP-DM methodology

Phase	Main tasks
1. Business understanding	Defining the business objective; translating the economic problem into a mathematical task; planning and risk assessment.
2. Data understanding	Collecting the initial data; exploring and checking data quality; identifying relationships between variables.
3. Data preparation	Cleaning and selecting data; handling missing values; generating new variables; formatting the dataset.
4. Modelling	Splitting into training/test sets; implementing and comparing techniques; assessing model accuracy.
5. Evaluation	Evaluating the model against the business objectives; deciding whether to deploy or to iterate the cycle.
6. Deployment	Planning deployment, monitoring and maintenance; preparing the final report; project review.

Source: *adapted from Chapman et al. (2000)*

The alternatives offer contextual advantages but also limitations: SEMMA is tightly bound to the SAS platform (Azevedo & Santos, 2008); KDD proposes 9 stages and is the academic precursor of data science; TDSP is optimised for Azure; Scrum and Kanban require adaptation for the unpredictable modelling phases. All the projects presented adopted CRISP-DM as their reference framework.

3. METHODOLOGY

This research adopts an applied multiple-case-study design combining qualitative analysis of project implementation with quantitative evaluation of model and business performance, with the aim of assessing how Data Mining techniques can generate measurable competitive advantages in the private sector of the Republic of Moldova. The choice of this method is justified by the applied nature of the research and by the need to analyse real projects implemented in different organisational contexts.

The research is based on three case studies carried out between 2018 and 2024 in organisations from the Republic of Moldova operating in distinct fields: debt collection, asset valuation for the financial and banking sector, and energy loss forecasting. The cases were selected to cover a variety of economic and operational problems, in order to demonstrate the cross-cutting applicability of Data Mining techniques in the private sector.

The research process was structured according to the CRISP-DM methodology, which served as the main framework for organising the analytical stages. Each project was thus analysed through the lens of the methodology's six phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment. This approach made it possible to link the organisations' economic objectives with the technical modelling and evaluation tasks.

In the first stage, the business problems were defined for each case analysed. For the debt collection scoring system, the objective was to estimate the probability that debtors would repay their debt. For asset valuation, the goal was the automated estimation of the market value of automobiles and apartments. For the energy project, the objective was to forecast technological consumption and electricity losses.

In the data understanding stage, the available sources, dataset structure, information quality, variable distributions and the presence of missing or outlier values were analysed. The data came from the operational sources of the partner organisations and were supplemented, where necessary, with relevant external data such as meteorological information, market characteristics or calendar indicators.

The data preparation stage included cleaning the datasets, removing or treating anomalous values, standardising formats, selecting relevant variables and constructing derived variables. Depending on the specifics of each project, transformation procedures were applied to numerical and categorical variables, along with noise-reduction techniques, to increase the robustness of the predictive models.

In the modelling stage, several Data Mining and Machine Learning techniques were tested, including decision trees, regression models, Random Forest, Gradient Boosting, autoregressive models and hybrid models. Algorithms were selected according to the nature of the problem — classification, regression or forecasting — as well as the requirements for interpretability, accuracy and operational deployability.

Model performance was assessed using metrics specific to each type of problem. For regression and forecasting models, measures such as R^2, RMSE, MAE and MAPE were analysed (Hyndman &

Koehler, 2006), while for the scoring model, classification accuracy, the stability of decision rules and the impact on operational indicators were monitored. The evaluation was not limited to the statistical performance of the models but also included an analysis of the economic relevance of the results obtained.

The solutions were validated by testing the models on data distinct from those used for training, in order to reduce the risk of overfitting and to verify their ability to generalise. Where the models were deployed in operational processes, validation was supplemented by monitoring post-deployment indicators such as fewer ineffective actions, faster valuation or improved forecast accuracy.

In the deployment stage, the model outputs were turned into operational tools: decision rules for prioritising collection actions, automated asset valuation models and forecasting systems integrated into energy analysis processes. The research thus aimed not only at obtaining high-performing models but also at integrating them into the organisations' day-to-day operations.

The methodology used therefore allows a comprehensive assessment of Data Mining projects from both a technical and a managerial perspective. This approach highlights the link between the quality of the analytical process, the choice of an appropriate methodology and the achievement of measurable competitive advantages in the private sector of the Republic of Moldova.

4. CASE STUDIES

4.1. Scoring system for debt collection

The first project was carried out in partnership with a debt collection company in the Republic of Moldova, with a portfolio of more than **300 clients** and experience of more than **1,450,000 collection cases** processed since its founding. The company's operations involved, every month, more than 40,000 active cases, 50,000 notices sent, 1,000 field visits to debtors' addresses and around 100,000 phone calls — a volume of activity that called for optimising resource allocation through advanced analytical methods.

The main objective of the project was to develop a predictive scoring model capable of estimating the probability that a debtor would voluntarily repay the debt within 6 months, in order to prioritise collection actions and reduce operational costs.

Table 3. Operational indicators of the partner company

40,000+	**50,000**	**1,000+**	**100,000+**
Cases processed monthly	*Notices sent monthly*	*Field visits monthly*	*Phone calls monthly*

The database used covered around **70,000 debtors** from the telecommunications segment. The model's dependent variable was repayment of the debt within 6 months (binary variable), and the predictors included the debt amount, the debtor's age, the age of the debt, the history of previous payments and other relevant behavioural indicators.

A significant methodological challenge identified in the project was *overfitting*: the model built on the entire database achieved an accuracy of around **80%**, but testing on new data showed a performance of only **55%**. This finding required revising the modelling strategy and adopting an approach based on *decision rules* derived from decision trees, with high probabilities of correct classification.

The final solution consisted in identifying decision nodes with high probabilities of non-payment, from which rules of the following type were formulated:

IF "Debt Amount" > 1589 AND "Debt Amount" < 3158 AND "AGE" > 46 THEN "NO PAY" WITH Probability = 0.89

The identified rules covered around **40%** of all records in the portfolio, representing **79%** of the total amount of debt managed. Deploying the model produced a measurable result: a 27% reduction in the number of unnecessary collection actions.

4.2. Automated market value estimation of assets

The second project focused on developing Machine Learning models for the automated estimation of the market value of two asset categories: **automobiles** and **apartments**. The project was implemented in partnership with the ***"VIA SCOPE" Credit History Bureau***, an institution that uses these tools to assess collateral and the creditworthiness of loan applicants.

For the **automobile** segment, data were collected and processed from seven distinct geographical sources: the Republic of Moldova, the USA, Germany, Poland, Russia, Romania and the United Kingdom. The final dataset included more than **3,400,000 records**, covering **80 makes** and **1,040 different models**, with **11 descriptive variables** per record (year of manufacture, mileage, engine capacity, fuel type, etc.).

For the **real estate** segment, **51,000 records** of apartment transactions in the Republic of Moldova were used, covering both the primary and secondary markets, with **12 descriptive parameters** (floor area, number of rooms, floor, geographical area, year of construction, etc.).

The modelling process involved implementing and systematically comparing **12 different models**, including linear regression, random forest, decision trees, gradient boosting, exponential smoothing and moving averages. The final model was selected on the basis of standard evaluation metrics (R^2, RMSE, MAE) on independent validation sets.

The performance of the selected models exceeded the initially set quality thresholds:

- Automobile value estimation: **$R^2 = 0.96$** — meaning that the model explains 96% of the variability in observed market prices;
- Apartment value estimation: **$R^2 = 0.90$** — a highly accurate model capable of generating reliable automated valuations for the collateral portfolios of partner financial institutions.

4.3. Forecasting system for technological consumption and losses — RED-Nord JSC

The third project focused on developing an intelligent system for forecasting technological consumption and electricity losses (TCL) in the distribution network of RED-Nord JSC — one of the largest electricity distribution operators in the Republic of Moldova.

The need for the project was driven by the scale of the losses recorded. The main objective of the system was to estimate TCL with high accuracy in order to support the identification and reduction of these losses, the optimisation of technological consumption and the strategic planning of infrastructure investment.

Architecturally, the system integrates data from internal sources (conventional and smart meters, SCADA systems) and external sources (weather data, calendar events, economic factors), processed through a structured pipeline that includes validation, cleaning, normalisation and outlier treatment, followed by storage in Data Lake and Data Warehouse structures. The modelling module uses three hierarchical levels of complexity: basic statistical analysis, statistical models for capturing trends and seasonality, and advanced Machine Learning techniques for complex and hybrid forecasts. The system integrates more than **twenty parameters**, including weather factors (temperature, precipitation, wind), special events and tariff variations.

The performance of the models developed significantly exceeded standard quality thresholds: the coefficient of determination reached **R^2 = 0.98**, and the mean absolute percentage error (MAPE) stood at **2.23%** — indicating that the model explains approximately 98% of the variability in the observed values, visible in the *forecast vs. actual* comparison charts generated by the system. The integrated dashboard provides hourly, daily, monthly and annual forecasts, simulation features based on weather and economic scenarios, and automated reports segmented by district and consumer category.

The projected economic impact of the system is substantial: a reduction of just **1%** in losses would generate savings of approximately **9,000 MWh** per year, with direct effects on operational efficiency and on the ability to report accurately to the National Agency for Energy Regulation (ANRE). In addition, automating data collection and analysis frees up significant human resources and reduces the risk of errors associated with manual data collection — identified as one of the major structural challenges facing the Moldovan energy system at the current stage of digitalisation.

5. RESULTS

A summary of the results of the three projects is presented in Table 4.

Table 4. Summary of Data Mining projects carried out in the Republic of Moldova

Project	Sector	Main method	Measurable result	Competitive advantage
Debt Collection Scoring	Telco	Decision trees, decision rules	−27% unnecessary actions	Resource prioritisation; lower operational costs
Automobile value estimation	Financial and banking	Random Forest, Gradient Boosting	R^2 = 0.96	Automated valuation vs. costly manual appraisal
Apartment value estimation	Financial and banking	Ensemble models	R^2 = 0.90	Consistency and scalability in collateral valuation
TCL forecasting — RED-Nord	Energy	Advanced ML	R^2 = 0.98; MAPE = 2.23%	Loss reduction: −1% = ~9,000 MWh/year

Source: *compiled by the authors based on the projects carried out*

The results obtained in the practical projects carried out in the Republic of Moldova confirm that the systematic application of Data Mining techniques, in line with the CRISP-DM methodology, generates concrete and measurable competitive advantages at the organisational level. The debt collection scoring model showed that optimising resource allocation through predictive methods significantly reduces operational costs, even with limited and imbalanced data. The automated asset valuation models — with coefficients of determination of 0.96 for automobiles and 0.90 for real estate — provide high predictive accuracy while offering clear advantages in scalability and consistency compared with fully manual valuation processes.

These results are particularly relevant for the Republic of Moldova, where the adoption of Data Mining techniques is still limited and their systematic use is associated mainly with organisations that have analytical skills, digital infrastructure and data-driven management practices.

Overall, the case studies presented provide empirical evidence in support of the central thesis of the paper: **Data Mining is a real and accessible competitive advantage for organisations in the Republic of Moldova**, provided that an appropriate methodological framework is chosen and the life cycle of the analytical project is thoroughly understood.

6. CONCLUSIONS

This paper demonstrates that Data Mining techniques are no longer merely an advanced analytical tool reserved for large companies or mature digital ecosystems, but a real source of competitive advantage for private organisations in the Republic of Moldova. The results of the three case studies confirm that, when applied within a rigorous methodological framework such as CRISP-DM, Data Mining methods can turn existing operational data into faster, more objective and more cost-effective decisions.

The main contribution of the paper lies in demonstrating the practical applicability of the CRISP-DM methodology in different economic contexts: debt collection, automated asset valuation and energy loss forecasting. In all three cases, data were treated not merely as informational support but as a strategic resource for optimising business processes. The scoring model made it possible to prioritise collection actions and reduce ineffective interventions; the automated asset valuation models helped increase the consistency and scalability of decision-making; and the energy forecasting system demonstrated the potential of Data Mining to support loss reduction and more efficient resource planning.

The empirical results presented confirm that the value of a Data Mining project should not be assessed solely by the statistical performance of its models, but also by its managerial and economic impact. Indicators such as R^2, MAPE, classification accuracy or reduced forecast error are important, but their relevance increases significantly when they translate into concrete operational decisions: lower costs, higher recovery rates, reduced losses, process automation and better decision quality.

The case studies suggest that organisational and analytical capabilities may represent important barriers to the wider adoption of Data Mining in the Republic of Moldova. In particular, insufficient analytical skills, a limited data-driven organisational culture, and managerial difficulties in translating economic problems into well-defined analytical tasks may constrain the effective use of available data (OECD, 2023). Many organisations already hold significant volumes of data, but these are used mainly for descriptive or reporting purposes rather than for prediction, optimisation and informing strategic decisions.

From a managerial perspective, the paper suggests that the successful implementation of Data Mining projects requires three essential conditions: a clear definition of the business problem, a structured process for data preparation and validation, and the integration of model outputs into the organisation's operational processes. Without this link between economic objectives and analytical solutions, even technically high-performing models may have limited impact on an organisation's competitiveness.

Data Mining should therefore be seen not merely as a technological component but as an instrument of organisational transformation. Its implementation supports the shift from decisions based predominantly on experience and intuition to decisions grounded in data, predictive models and quantifiable assessments of risks and opportunities. In this respect, CRISP-DM offers a robust, flexible and accessible framework for organisations seeking to build in-house analytical capabilities and make better use of the data they hold.

Future research could extend the analysis to other sectors of the national economy, such as retail, agriculture, healthcare, transport and public services. It would also be useful to investigate the relationship between organisations' analytical maturity, the quality of available data and the performance of Data Mining projects. Another important research direction is the development of models for assessing the return on investment in Data Science projects, to help managers make better-founded decisions on digitalisation and process automation.

In conclusion, the experience analysed confirms that Data Mining can become an important driver of modernisation and competitiveness in the private sector of the Republic of Moldova. Organisations that succeed in investing in analytical skills, data governance and clear implementation methodologies will be able to turn data into economic value, competitive advantage and better-founded strategic decisions. This conclusion is supported by the concrete results of the case studies presented in this article, including fewer unnecessary collection actions, the high accuracy of the valuation models and the performance of the energy forecasting system.

REFERENCES

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: A parallel overview. In *Proceedings of the IADIS European Conference on Data Mining 2008*. IADIS.

Bramer, M. (2016). *Principles of data mining* (3rd ed.). Springer.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc. https://mineracaodedados.wordpress.com/wp-content/uploads/2012/12/crisp-dm-1-0.pdf

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*.

Mariscal, G., Marbán, Ó., & Fernández, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., & Flach, P. (2021). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*.

OECD. (2023). *Promoting digital business skills in the Republic of Moldova*. OECD Publishing. https://doi.org/10.1787/9d35a60f-en

Piatetsky, G. (2014, October). *CRISP-DM, still the top methodology for analytics, data mining, or data science projects*. KDnuggets. https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Rohan, R. K. (2023). *Understanding the differences: AI, data science, data mining, machine learning, and deep learning*. Medium.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*.

Shimaoka, A. M., Cordeiro Ferreira, R., & Goldman, A. (2025). Leveraging XP and CRISP-DM for Agile Data Science Projects, https://arxiv.org/pdf/2505.21603