Machine Learning Empowerment in Industry 4.0 – Case Study for Micro and Small Enterprises in Romania

Flavia BOGOEVICI*
Bucharest University of Economic Studies, Bucharest, Romania
**Corresponding author, flavia.bogoevici@yahoo.com*

Octavia ALBU
Bucharest University of Economic Studies, Bucharest, Romania
albu.a.octavia@gmail.com

Ruxandra DUȚĂ
Bucharest University of Economic Studies, Bucharest, Romania
ruxandra2013@gmail.com

Camelia CHITCA
Bucharest University of Economic Studies, Bucharest, Romania
camelia_chitca@yahoo.com

Abstract. *In the world in which technology quickly integrates in our daily lives, businesses that incorporate digital innovation throughout their organizational culture, spanning from top-level executives to low-level employees are prone to emerge as industry frontrunners. Supported by Machine Learning, which stands out as a pivotal revolutionary tool, companies can enhance their productivity and operational efficiencies by incorporating remarkable automation capabilities, error reduction, superior predictive analysis, together with gaining valuable insights into future trends. The paper confers an overview of Machine Learning's capabilities, developed types, provided solutions and built architecture, through a conceptual structure. The paper elaborates these crucial concepts, offering a precise perspective on the topic and adopts a descriptive approach, elucidating the provided terminologies and ideas by referencing the related literature. The paper highlights in the initial part the outcomes resulting from the key advantages of Machine Learning and its impact on organizations, the path towards realizing substantial value through these digital advancements, emphasizing the priority organizations assign to cultivate their digital potential. The research performed in the second part of the paper aims at analyzing the progress of Romanian micro and small enterprises with implemented Machine Learning solutions, with detailed metrics and comprising k-means clustering, having the following objectives: automating repetitive tasks, improving planning and forecasting, increasing net profit, effortlessly discovering new patterns from large, diverse data models.*

Keywords: Machine Learning, Clustering, Industry 4.0, Prediction, Optimization

Introduction

The trending concept of Industry 4.0 holds the promise of remarkable outcomes for industries, while bringing about profound changes in various aspects of organizations. These changes encompass the establishment of business models, the entire production process and extend to the point when the customer receives the product. Despite being a newly emerging concept with a broad range of definitions, there is a current lack of knowledge regarding specific steps and only a few scenarios available for guiding the transition to Industry 4.0. A new set of technologies is

DOI: 10.2478/picbe-2024-0274

emerging, indicating significant societal changes and justifying the preliminary designation as a fourth industrial revolution. Industry 4.0 is a term that is highly popular but still lacks a precise definition. Also, it serves as an umbrella term for advancements in manufacturing technologies, especially referring to Machine Learning, Artificial Intelligence, Robotic Process Automation, cloud computing, Internet of Things, predictive maintenance. (Suleiman et al., 2022)

Literature review

Understanding machine learning

Machine learning (or ML) is a subset of artificial intelligence that uses algorithms to allow systems to learn and improve from experience or data without explicit programming. It focuses on creating models capable of generalizing trends and making data-driven predictions or decisions.

Machine Learning can be conceptualized as the process through which computer algorithms improve their performance at a given task with increased exposure to data, without being explicitly programmed for every possible scenario. (Kanade, 2022)

This concept can be illustrated through a simplified example aimed at distinguishing between two categories: cats and dogs. Ashan considers the scenario in which a robot is to be trained to tell pictures of cats and dogs apart. Initially, the robot might make mistakes because it is just starting to learn and lacks the ability to accurately classify the images. However, as it shows more and more pictures, it starts noticing details that make cats different from dogs, like the shape of their ears, the size of their tails, or their fur. Over time, and with the analysis of a sufficiently large and diverse dataset, the algorithm develops a model that encapsulates the *learned features and patterns* associated with cats and dogs. This model enables the algorithm to classify new, unseen images with a degree of accuracy that improves as the size and quality of the training dataset increases. (Ahsan, 2023)

The example underscores the essence of machine learning: *enabling computer algorithms to autonomously learn from and make predictions based on data.* Such algorithms are not limited to image recognition, but extend to a myriad of applications across various domains, including but not limited to, healthcare diagnostics, financial market analysis and natural language processing. The key advantage of machine learning lies in its ability to adapt and refine its predictions over time, thereby offering solutions that are both scalable and applicable to complex, real-world problems.

Figure 1. Machine learning process

Source: Authors' own creation, based on Ahsan, 2023.

Hidden in plain sight, machine learning, in one form or another, lies behind Netflix's recommendation system (Steck et al., 2021), your home's or phone's virtual assistant (Kepuska & Bohouta, 2018), or even your social media feed (Brown, 2021). However, these commonplace applications serve as a foundation for delving into the diverse trends, functions and use cases currently influencing the field of machine learning.

In the following table, the key characteristics of supervised, unsupervised and semi-supervised machine learning types are outlined and compared to provide a clear understanding of their distinct methodologies and applications:

Table 1. Machine learning types

Subcategory	***Description***	***Main tasks***	***Example***
Unsupervised	Uncovers hidden patterns in unlabeled data through clustering, association and dimensionality reduction, without needing human intervention to guide the analysis.	1. Clustering: data mining technique for grouping unlabeled data based on their similarities or differences 2. Association: method that uses different rules to find relationships between variables in a given dataset 3. Dimensionality reduction: a learning technique used when the number of features in a dataset is too high. It reduces the number of data inputs to a manageable size while also preserving the data integrity	Using the K-means algorithm to group customers into clusters based on their purchasing behavior for market segmentation (e.g. market basket analysis - customers who bought this item also bought something else)
Supervised	The algorithms are trained on labeled data to accurately classify information or predict outcomes.	1. Classification: categorizing data into predefined groups, using algorithms like linear classifiers, support vector machines (SVMs) or decision trees 2. Regression: predicts continuous outcomes by understanding the relationship between dependent and independent variables in a dataset	Identifying whether an email is spam or not, or predicting the selling price of a house based on its size, location and other features.
Semi-supervised	Combines both labeled and unlabeled data in its training set, making it particularly effective for tasks like improving the accuracy of medical image analysis with limited labeled examples.	1. Classification: categorizing data into predefined groups with the help of both labeled and unlabeled data 2. Regression: predict continuous outcomes, but they benefit from the inclusion of unlabeled data in training.	Improving facial recognition systems. A small amount of facial images are labeled with the person's identity and the system uses a larger amount of unlabeled facial images to enhance its ability to accurately recognize and classify new faces.

Source: Authors' own creation, based on Delua (2021).

ML functionalities across industries

It is well known that Machine Learning (ML) has a disruptive impact across a wide range of industries, transforming how businesses operate, make decisions and connect with customers. From healthcare predictive diagnoses to finance risk assessment, retail tailored shopping experiences and industrial predictive maintenance, ML's possibilities are extensive and diverse.

Through the most important ML solutions there can be counted: customer analytics, predictive analytics/modeling, anomaly detection, NLP, computer vision.

Content sentiment analysis

Understanding sentiment analysis is foundational for leveraging AI in *customer feedback analysis*. This process automatically analyzes large text datasets, such as customer reviews, to detect the represented feelings, meaning positive or negative ones. The goal is to understand general customer attitudes, which will help firms identify preferences, areas for improvement, and make informed, data-driven decisions. As e-commerce grows, AI-powered sentiment analysis becomes increasingly important, enabling for efficient and automated analysis of client comments. This automated technique *speeds up insight extraction by substituting time-consuming manual reviews.*

The use of ML for sentiment analysis offers several benefits, making it an increasingly favored tool for analyzing customer feedback. The benefits are highlighted in the "AIContentfy" article from 2023, which include:

1. Speed and efficiency: the use of ML algorithms contributes significantly to the reduction of time in processing large text datasets. This quick analysis allows companies to promptly identify patterns and trends in customer sentiment, enabling real-time responses to emerging issues.
2. High potential of accuracy: the ability of ML models to learn from large datasets, recognize patterns, adapt to new information, handle ambiguity, generalize across diverse contexts and reduce bias collectively contributes to the improved accuracy observed in sentiment analysis tasks.
3. Consistency: the standard approach, objectivity, automation, scalability and elimination of human mistakes all help at improving the overall consistency of sentiment analysis powered by machine learning.
4. Scalability: the sentiment analysis allows organizations to handle large volumes of data, ensures real-time processing, accommodates global operations, adapts to growth, automates workflows and maintains consistent performance across different data sizes.
5. Trends: ML facilitates the identification of trends in sentiment analysis by recognizing patterns, conducting temporal analysis, assessing product/service performance, comparing sentiments across segments, identifying influencing factors, enabling predictive analysis and establishing a feedback loop for continuous improvement.

This being said, sentiment analysis using a machine learning approach still has some possible *challenges and limitations* ahead of its great possible future. The key considerations, as told by the AIContentfy team, could be:

1. Accuracy Limitations: the accuracy of AI-powered sentiment analysis is closely linked to the quality of the training data. Biased or unrepresentative data can lead to inaccuracies, as the algorithm may not generalize well to diverse or unseen inputs.
2. Multilingual analysis: variations in grammar, expressions and cultural nuances pose challenges, making it challenging for ML algorithms to provide accurate analyses across diverse languages.
3. Privacy concerns: when dealing with text containing sensitive or personal information, some ethical concerns arise from how this data is collected, stored and used, emphasizing the need for strong privacy protections.

In summary, understanding sentiment analysis is essential for harnessing the power of ML in customer feedback analysis, as highlighted by the AIContentfy team (2023). This process, crucial for discerning positive or negative sentiments in large text datasets like customer reviews, aids in gauging general customer attitudes.

Price optimization

ML models, as those developed by N-iX company, play a crucial role in price optimization by using both labeled and unlabeled data to predict demand and set optimal prices.

Techniques such as regression models, sequence models like LSTM and time-series models including ARIMA, along with tree-based models like Random Forest and XGBoost, are employed for their ability to forecast customer demand and adjust pricing strategies accordingly. Furthermore, the price elasticity of demand models and linear programming optimization techniques aid in understanding demand changes in response to price alterations and in finding optimal pricing strategies to maximize revenue. (Fedynyshyn, 2023)

A step-by-step implementation process involves *data collection, cleaning, feature engineering, model selection and training, validation and testing, deployment and continuous monitoring and updating* to ensure the effectiveness of the pricing optimization system. Real-life applications include automating tariff recalculation for a UK energy supplier, significantly reducing manual update time and implementing a product pricing platform for a top US retail enterprise, enhancing operational accuracy and data analytics capabilities.

Despite the potential of ML for price optimization, *challenges* like historical data scarcity, cross-pricing impact and supply chain volatility must be addressed. Solutions include employing reinforcement learning, integrating cross-pricing effects and incorporating real-time supply chain data into models to enhance adaptability and accuracy in dynamic market conditions.

Recommender Systems

Recommender systems have emerged as a significant technological advancement, particularly highlighted by initiatives like Netflix's Prize in 2006, aimed at enhancing its recommendation accuracy. These systems, utilizing vast datasets, play a pivotal role in various industries, including entertainment and banking, by personalizing user experiences based on data analysis. For instance, Netflix leverages user data to recommend content, significantly influencing viewer activity, while TikTok optimizes user engagement through personalized video recommendations using reinforcement learning. (Bouvier, 2021)

It is also known that multiple industries benefit from recommender systems, for example the **banking industry**, which offers tailored financial products and services, thereby increasing both direct and indirect revenues. These systems can suggest products like insurance bundles to customers based on their purchase history or provide personalized savings advice, enhancing customer satisfaction and loyalty. Furthermore, the implementation of recommender systems in banking, demonstrated by institutions like Santander and Emirates NBD's Liv, showcases their capability to *predict customer needs, streamline administrative tasks and offer bespoke financial advice with minimal human intervention.* The future of banking is expected to be dominated by *personalization*, with recommender systems at the forefront, enabling banks to deliver more tailored and efficient services. This trend towards customization, supported by technologies like NLP, not only improves customer interaction with banks, but also significantly boosts engagement and revenue. Moreover, the application of these systems extends to loan customization and robo-advisory services, indicating a shift towards more individualized and need-based financial solutions. Despite the increasing reliance on technology, the human element remains crucial in banking, emphasizing the importance of a balanced approach that combines technological efficiency with human empathy and trust. (Bouvier, 2021)

Another important industry in which machine learning plays a crucial role is the *pharmaceutical industry*. The Drug Recommender System employs advanced machine learning techniques, including sentiment analysis and drug recommendation algorithms, to provide valuable insights into drug prescriptions and treatments. By analyzing user sentiments and emotions in text data, the system categorizes attitudes, offering information on medication effectiveness, adverse effects and patient satisfaction. The integrated approach aims to optimize medicine selection, improve patient outcomes and reduce adverse reactions. Through the evaluation of a large dataset, the system demonstrates its ability to make accurate and personalized recommendations, potentially revolutionizing healthcare by empowering patients and doctors to make informed decisions. (Lavanya & Praveen, 2023)

The following diagram presents the system architecture steps for machine learning models:

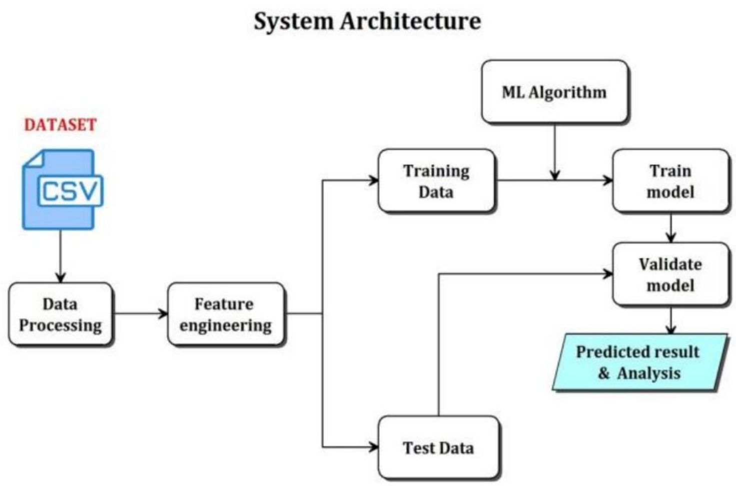

Figure 2. System architecture steps for Machine Learning

Source: Ankitha & Prakash (2022).

A. Data Gathering and Preprocessing: by data gathering there are collected signals representing real-world situations, which are then converted into electronic integer values, while data preprocessing transforms unclean data into clean datasets.

B. Feature Selection and Data Preparation: generating features from input data improves the prediction capacity of machine learning models. Feature engineering is a key skill that significantly influences the success of a machine learning model. In this sense, data classification involves grouping and categorizing data based on specific characteristics, either numerical or attribute-based.

C. Model Construction and Model Training: training a machine learning model involves providing a learning algorithm with a training set to learn from. The ”Machine-Learning model" outcome is generated during this training process. The training data should include the correct solution (goal or target attribute) and the learning method constructs a model by identifying patterns in the training data that relate the input data characteristics to the target.

D. Model Verification and Outcome Evaluation: The model is used to analyze new input during testing, with separate samples for training and testing data. The machine learning technique is designed for effective performance and generalization to new data in both sets. Real-time data is then input for prediction once the model is evaluated. After making a forecast, the result is examined for key information. (Ankitha & Prakash, 2022)

As it can be observed in the following graph, the Machine Learning market is projected to reach US $204.30 billion this year. Furthermore, the market size is expected to have a CAGR of 17.15% during the 2024-2030 and a market volume of US$528.10 billion by 2030. At a global scale, the largest market size will be in the United States (US $70.63 billion in 2024).

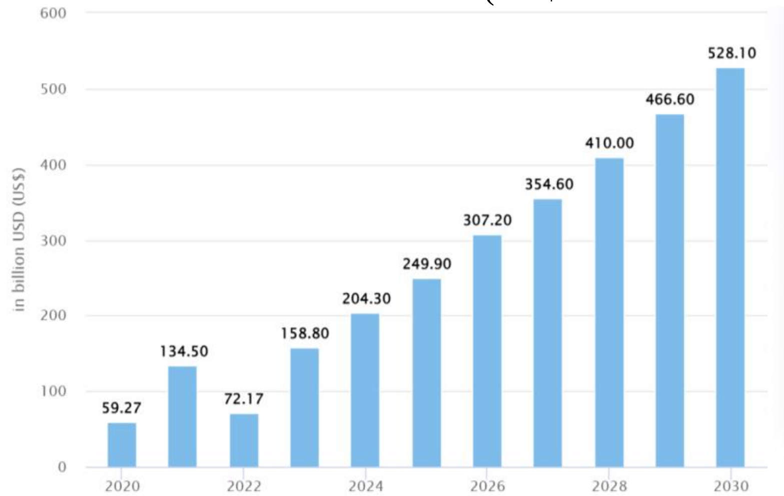

Figure 2. Machine Learning market size

Source: Statista (2024).

Methodology

The main purpose of the present research is to systematically assess the progress of *micro and small enterprises (MSEs) that have been engaged in Machine Learning activities in Romania during the 2020-2022 period.*

Being a descriptive research, it references the related literature to define the conferred terminologies and the attained conclusions. Both *quantitative* and *qualitative* data has been collected, more specific *primary data* (collected for the first time by researchers, in this case from the two platforms where the MSEs are listed) and *secondary data* (already collected, referring to the explanatory notes from journals and press articles).

MSE is a term used to describe businesses that maintain a range of employees, revenues or assets that fall below certain thresholds, distinguishing them from larger corporations (Bigliardi et al., 2011). These thresholds can vary by country but in general, they aim to categorize companies according to their size and operational capacity. MSEs are recognized for their crucial role in the economy, acting as vital engines of growth, innovation and employment.

In the rapidly evolving landscape of machine learning and artificial intelligence, understanding the vitality and potential of MSEs within this domain requires a nuanced and holistic approach. The dynamic nature of this sector, coupled with the unique challenges and opportunities it presents, necessitates the consideration of various factors that influence a company's performance and prospects. By examining the *geographical location, primary domain of activity, annual revenue, profits, average number of employees and year of establishment*, valuable insights into the operational health, strategic positioning and innovation capacity of these entities are to be gained. These factors collectively offer a comprehensive framework for understanding the complexities of the machine learning industry in Romania. They not only reflect the immediate economic and competitive standing of these MSEs, but also their potential for sustainable growth and adaptation in a technology-driven market (Bigliardi et al., 2011).

In Romania, SMEs are defined based on criteria established by the European Commision, which classifies companies into categories according to their number of employees and their annual turnover, specifically (*Microîntreprinderile și Întreprinderile Mici și Mijlocii: Definiția și Sfera De Cuprindere*, 2016):

- Micro-enterprises: companies with fewer than 10 employees and an annual turnover of less than 2 million euros;
- Small enterprises: companies with fewer than 50 employees and an annual turnover of less than 10 million euros;
- Medium-sized enterprises: companies with fewer than 250 employees and an annual turnover of less than 43 million euros.

Data extraction

For the purpose of this research, the identification and selection of MSEs operating within the machine learning domain in Romania were conducted using *Crunchbase*, a comprehensive platform for business information about private and public companies (What Is Crunchbase and How Does It Work?, 2022). By setting the search parameters to focus on *Romania and the specific field of ML, a subset consisting of 58 MSEs was considered.* This approach ensured a selection of both targeted and relevant companies.

The initial identification process utilized Crunchbase's advanced search capabilities to filter companies by geographical location (Romania) and primary domain of activity (Machine Learning), thereby compiling a preliminary list of relevant enterprises. Subsequently, to ensure the accuracy and reliability of the data pertaining to these companies, a *cross-verification* was conducted using the *Risco* platform, which served as the primary source for retrieving company-specific information. This included net profit, annual turnover, average number of employees, the NACE code and the year of establishment. The cross-verification process not only affirmed the validity of the initial findings, but also enriched the dataset with essential financial and operational metrics.

In the analysis of the companies, several key variables were focused on in order to gain insight into their performance, as follows:

1. Company name: the primary identifier for each enterprise;
2. Headquarters county: the geographical location of the company's headquarters, that provides insights into regional economic activity;
3. Annual turnover for 2020, 2021 and 2022: refers to the total amount of revenue (or sales) an enterprise generates over a year. It serves as a measure of the company's ability to sell goods and services within a set timeframe, indicating the scale of a business's operations and its effectiveness in generating revenue from its core activities (Lenglet, 2023);
4. Net profit for 2020, 2021 and 2022: refers to the amount of revenue that remains after a company has deducted all its costs, expenses and any other charges from its total revenue. It is a crucial indicator of a company's financial performance and profitability (Davis, 2017);
5. Average number of employees for 2020, 2021 and 2022: provides information about a company's size and scalability;
6. NACE (Nomenclature of Economic Activities) code: represents the company's primary area of activity according to the European classification system;
7. Year of establishment: offers information about an enterprise's maturity, resilience and experience within its industry.

In the context of this research, NACE codes play a crucial role in understanding and categorizing the economic activities of the companies analyzed. CAEN stands for "Clasificarea Activităților din Economia Națională", which translates to the Classification of Activities in the National Economy. This system is the Romanian equivalent of the NACE codes used in the European Union. Within the database utilized for this research, specific The NACE codes correspond to the primary economic activities of the analyzed companies. These codes not only help in identifying the industry sectors these companies belong to but also allow for sector-specific analyses, such as comparing growth rates, profitability, and employment trends within and across sectors.

The following table will explain the specific NACE codes found in the database:

Table 2. NACE codes description

NACE code	Description	Industry
3320	Manufacture of musical instruments	Manufacturing
4321	Electrical installation	Construction
4791	Retail sale via mail order houses or via Internet	Wholesale and retail trade
5630	Camping grounds (including grounds for camping vehicles), and campsites	Hotels and restaurants
5829	Other software publishing	Information and communication
5911	Motion picture, video and television programme production activities	Information and communication
6201	Computer programming activities	Information and communication
6202	Computer consultancy activities	Information and communication
6209	Other information technology and computer service activities	Information and communication
6311	Data processing, hosting and related activities	Information and communication
6312	Web portals	Information and communication
7022	Business and other management consultancy activities	Professional, scientific and technical activities
7120	Food quality testing and analysis	Professional, scientific and technical activities
7219	Other research and experimental development on natural sciences and engineering	Professional, scientific and technical activities
7430	Translation and interpretation activities	Professional, scientific and technical activities
8559	Educational support activities	Education

Source: (Domenii CAEN - Lista Completa Coduri CAEN Cu Echivalenta NACE Si SIC, 2024).

Results and discussions

The following section aims at presenting the situation and progress of the chosen companies which have implemented Machine Learning solutions, showing metrics and highlighting the efficiency of this technology by their increase in profit.

In the figure below all the NACE codes that can be selected for a specific county in order to visualize the situation of its corresponding average profit, turnover, minimum and maximum number of employees during the 2020-2022 period are displayed.

NACE Code (All)

Row Labels	Average of Net Profit - Ron 2020	Average of Net Profit - Ron 2021	Average of Net Profit - Ron 2022	Average of Turnover 2020	Average of Turnover 2021	Average of Turnover 2022	Max of no. of employees (on avg-2020)	Min of No. of employees (on avg-2020)	Max of No. of employees (on avg-2021)	Min of No. of employees (on avg-2021)	Max of No. of employees (on avg-2022)	Min of No. of employees (on avg-2022)
MURES	(1,625,143.00)	(1,595,355.00)	(676,873.00)	1,648,648.00	1,436,641	1,578,550	4	4	3	3	3	3
PRAHOVA	(1,089,364.00)	(2,446,065.50)	(6,397,369.25)	1,401,069.25	3,316,818	8,367,156	28	4	43	7	68	3
BUZAU	(248,487.00)	642,795.00	1,062,041.00	7,116,217.00	9,411,579	6,694,392	12	12	9	9	9	9
HARGHITA	(12,356.00)	(12,356.00)	(12,356.00)	-	-	-	0	0	0	0	0	0
ALBA		46,876.00	448,448.00		-	2,450			2	2	8	8
ARGES		(2,699.00)	(27,488.00)		134,319	149,739			3	3	3	3
GIURGIU		(42,252.00)	1,769,009.00		-	2,214,199			2	2	3	3
OLT	3,162.00	(25,664.00)	(835,086.00)	4,385.00	23,572	215,396	0	0	1	1	1	1
ILFOV	23,546.00	12,430.00	(38,668.00)	93,297.00	112,015	35,890	0	0	1	1	1	1
HUNEDOARA	23,685.00	(1,231.00)	(3,215.00)	32,297.00	6,010	-	0	0	0	0	0	0
IASI	60,468.33	438,644.75	678,365.75	2,355,936.33	2,372,096	3,890,016	26	2	28	0	37	0
TELEORMAN	101,445.00	488,716.00	1,061,714.00	1,606,844.00	2,854,745	4,016,779	8	8	7	7	11	11
BRASOV	409,534.33	1,092,238.67	506,639.75	6,160,161.33	8,728,483	6,620,578	107	1	120	2	114	1
CLUJ	458,984.78	1,113,882.40	2,347,348.33	3,536,765.11	6,675,725	11,445,615	68	0	84	0	103	0
MARAMURES	821,392.00	1,866,851.00	1,367,010.00	1,273,433.00	2,846,430	3,648,886	16	16	13	13	13	13
TIMIS	881,476.50	727,672.33	824,768.67	4,134,020.50	3,215,023	3,786,441	28	1	28	1	31	1
GALATI	1,264,726.00	723,082.00	473,976.00	3,112,394.00	2,525,879	2,908,045	18	18	12	12	13	13
BUCURESTI	2,331,292.84	3,138,995.29	5,369,524.48	15,582,694.89	20,339,019	29,046,703	195	0	211	0	235	0
Grand Total	**971,283.71**	**1,343,666.11**	**2,107,008.98**	**7,800,208.47**	**10,031,824.61**	**14,358,130.24**	**195**	**0**	**211**	**0**	**235**	**0**

Figure 3. Metrics of the Romanian MSEs with implemented Machine Learning, classified by county

Source: Authors' own creation.

The figures suggest economic disparities among regions, with some showing robust growth in both profits and employee strength, while others face significant losses. For instance, Bucuresti exhibits the highest net profits and turnovers, indicating a thriving economic status. Conversely, regions like Mures and Prahova show considerable financial challenges.

The below graph highlights that for MSEs in Bucharest with NACE code 6201 (Computer programming activities in the IT industry) the turnover has increased over the last years, having an average value of 14 million lei in 2022.

NACE Code	6201
HQ County	BUCURESTI

Turnover 2020 - Average	Turnover 2021 - Average	Turnover 2022 - Average
8,802,622	12,347,246	13,997,503

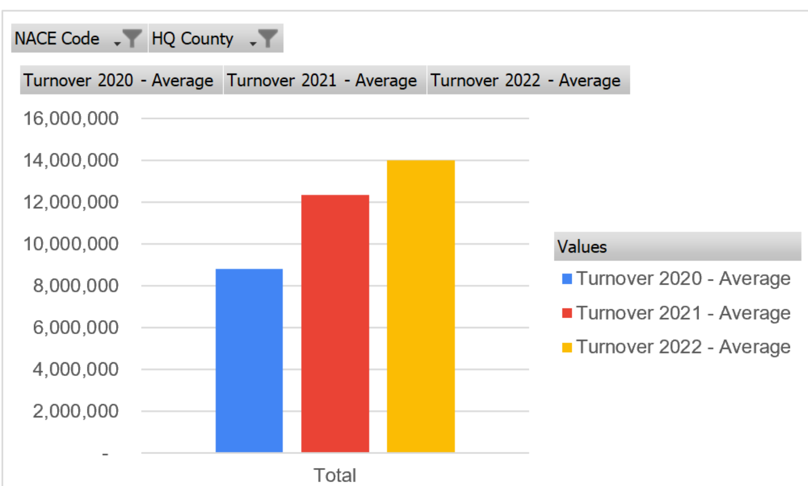

Figure 4. Turnover increase for MSEs in Bucharest with Machine Learning implemented, 2020-2022

Source: Authors' own creation.

DOI: 10.2478/picbe-2024-0274, pp. 3357-3373, ISSN 2558-9652 |
Proceedings of the 18th International Conference on Business Excellence 2024

The professional, scientific and technical activities industry with 7219 code - other research and experimental development on natural sciences and engineering recorded increasing values in net profit from the beginning of 2020, at national level, in value of 772,570 lei in 2022, as it can be observed:

NACE Code	7219
HQ County	(All)

Net Profit 2020 - Average	Net Profit 2021 - Average	Net Profit 2022 - Average
(12,716)	208,697	772,570

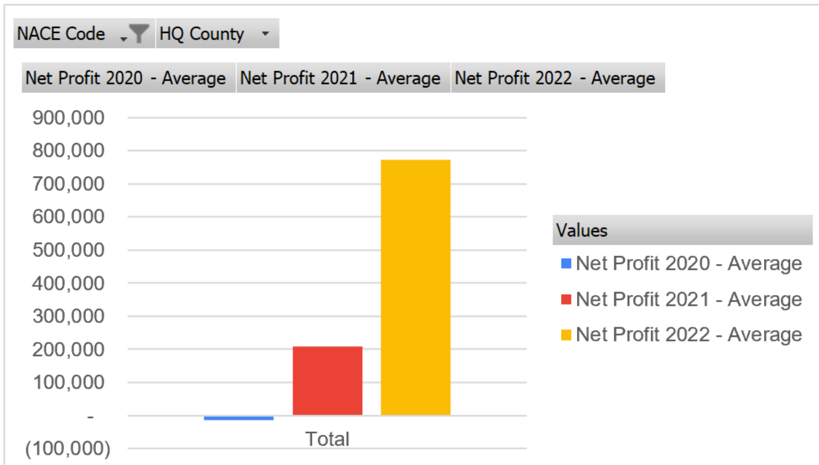

Figure 4. Net profit increase for MSEs in Romania with Machine Learning implemented, 2020-2022

Source: Authors' own creation.

Following, the companies with the highest 2022 turnover, Conectys being the leader, with a value of 400,810,866 lei, followed by R Systems Computaris Europe, with 112,768,733 lei, can be observed:

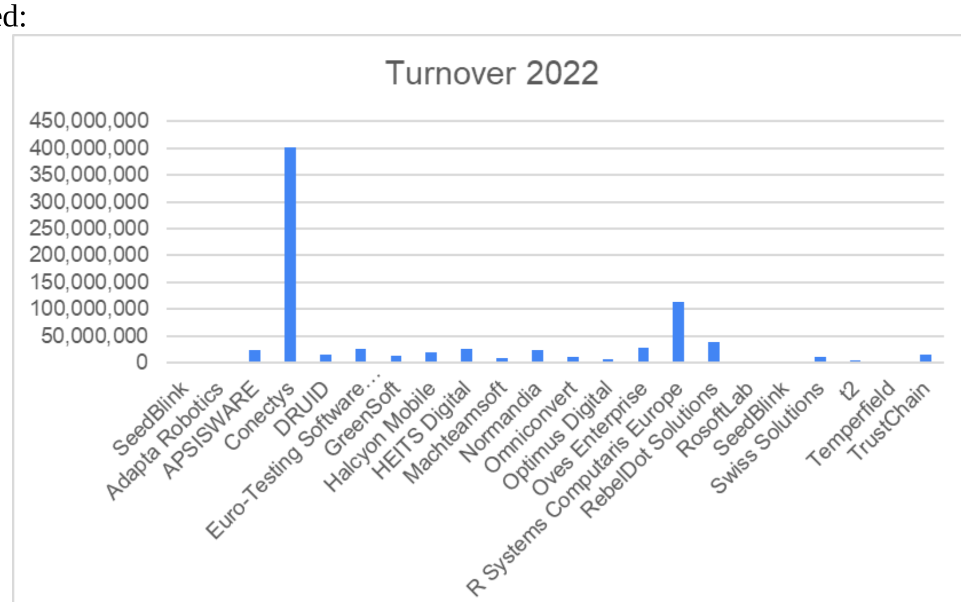

Figure 5. Highest turnover for 2022

Source: Authors' own creation.

Further on, a *heat map* (a graphical representation of data where chosen individual values are represented as colors) is desplayed. For establishing the criteria for the values, data was divided in four quarters, sorted ascending to each quarter's value and set according to the minimum, maximum and median values. It is worth mentioning that the average number of employees is a whole number, while the net profit and turnover are expressed in lei, adjusted to inflation.

Company Name	Turnover 2020	Turnover 2021	Turnover 2022	Avg. Employee No. 2020	Avg. Employee No. 2021	Avg. Employee No. 2022	Net Profit 2020	Net Profit 2021	Net Profit 2022
SeedBlink	1	1	1	1	1	1	2	3	2
4E Software	1	1	1	1	1	1	2	3	4
Adapta Robotics	1	1	1	1	1	1	2	4	4
APSISWARE	3	2	3	3	2	2	3	3	4
Artificial Intelligence Expert	1	1	1	1	1	1	2	3	4
Bitcom IOT	1	1	1	1	1	1	2	3	4
BufSecurity Systems	1	1	1	1	1	1	2	3	4
ClientZen	1	1	1	1	1	1	2	3	4
CMAT Industry	1	1	1	1	1	1	2	3	4
Dataworks Research SRL	1	1	1	1	1	1	2	2	3
DRUID	1	1	2	2	2	3	1	1	1
Euro-Testing Software Solutions	4	4	3	3	3	2	4	4	4
Footprints AI	1	1	1	1	1	1	2	3	4
Gray Matter Software	1	1	1	1	1	1	2	3	4
GreenSoft	1	1	2	1	1	2	2	3	4
GuidefAi	1	1	1	1	1	1	2	3	4
Halcyon Mobile	2	2	2	3	2	2	3	4	4
Happy Recruiter	1	1	1	1	1	1	1	3	3
HEITS Digital	1	2	3	1	1	1	3	4	4
Helios Vision	1	1	1	1	1	1	2	3	4
HiringDNA	1	1	1	1	1	1	2	3	4
Holisun	1	1	1	1	1	1	3	4	4
Machteamsoft	2	2	1	2	1	1	3	4	4
Marionette Studio	1	1	1	1	1	1	2	4	4
Atom Industries	1	1	1	1	1	1	2	3	4
Nenos Software	1	1	1	1	1	1	3	4	4
Normandia	2	3	3	4	4	4	3	4	4
Noteb	1	1	1	1	1	1	2	3	4
Omniconvert	1	2	2	2	2	2	2	3	3
Optimus Digital	1	1	1	1	1	1	2	4	4
Oves Enterprise	1	2	3	1	1	1	3	4	4
Potri Tehnology	1	1	1	1	1	1	2	3	4
QTEAM SOFTWARE SOLUTIONS	1	1	1	1	1	1	2	3	4
RebelDot Solutions	2	3	4	2	3	4	3	4	4
RECYCLLUX	1	1	1	1	1	1	2	3	4
REFRESH YOUR MIND CONSULTING	1	1	1	1	1	1	2	3	4
RepsMate	1	1	1	1	1	1	2	3	4
RosoftLab	1	1	1	1	1	1	2	3	4
Sag Services Provider	1	1	1	1	1	1	3	4	4
SecurifAI	1	1	1	1	1	1	2	3	3
SeedBlink	1	1	1	1	1	1	2	3	2
Sensix Environment	1	1	1	1	1	1	2	3	4
Skillview IO	1	1	1	1	1	1	2	3	4
Smart Touch Technologies	1	1	1	1	1	1	2	3	4
STEEPSOFT	1	1	1	1	1	1	2	3	4
Strongbytes	1	1	1	1	1	1	2	3	4
Swiss Solutions	2	1	2	2	1	2	3	4	4
SYNAPTIQ TECHNOLOGIES	1	1	1	1	1	1	2	3	4
t2	1	1	1	1	1	1	2	3	4
Tembo Island café	1	1	1	1	1	1	2	3	4
Temperfield	1	1	1	1	1	1	3	3	4
trendometrics	1	1	1	1	1	1	3	4	4
TrustChain	1	1	2	1	1	1	2	4	4
Vatis Tech	1	1	1	1	1	1	2	3	3
Woof Diet	1	1	1	1	1	1	2	3	4
Zetta Cloud	1	1	1	1	1	1	3	3	4

Figure 6. Heatmap of Romanian MSEs that implement ML

Source: Authors' own creation.

K-means clustering
Data preparation

A critical step in the data preparation phase of the analysis involved the identification and handling of outliers within the dataset, which was accomplished through the utilization of the *Z-score method*.

The Z-score, a statistical measure, quantifies the number of standard deviations a data point situates itself from the mean of a data set (Zhang et al., 2014). It represents a powerful tool in identifying outliers, as it provides a standardized way to assess the variability of data points. By calculating the Z-score for each data point in variables such as net profit, annual turnover and average number of employees, data points that fell beyond a predefined threshold (commonly set at Z-scores greater than |3|) were identified as outliers. These outliers represent values that are unusually high or low in comparison to the overall distribution of the data and may indicate anomalies, measurement errors, or genuinely exceptional cases within the dataset. The identification of outliers through Z-scores is pivotal in data preparation, as it ensures the robustness and reliability of subsequent analyses by minimizing the influence of extreme values that could skew results.

Clustering

K-means clustering is a popular unsupervised learning algorithm used to partition a dataset into K distinct, non-overlapping clusters. It aims to categorize data points into clusters in such a way that similar data points are grouped together, thereby revealing inherent patterns or groupings within the data. The "K" in K-means represents the number of clusters to be identified, and it must be specified by the user at the outset (Burkardt, 2009).

The process, as described by Burkardt, begins with selecting initial centroids, either randomly or through advanced strategies. These centroids act as the provisional centers for the clusters. The algorithm then iterates through a two-step process until it achieves stability. Initially, data points are assigned to the closest centroid, typically determined by the Euclidean distance, effectively forming K preliminary clusters.

This iterative process continues until the centroids stabilize and no longer change significantly, indicating that the clusters are as internally consistent yet as distinct from each other as possible.

K-means clustering's appeal lies in its simplicity and efficiency, making it suitable for a wide array of applications, from market segmentation to organizing computing clusters. However, its success hinges on the initial selection of centroids and the chosen number of clusters, K. The Elbow Method is a technique used to determine an optimal K by balancing the need for distinct clusters against minimizing the variance within each cluster. This method and the algorithm's iterative nature allow for the adaptive discovery of inherent groupings within large datasets, showcasing K-means' versatility and power in uncovering data insights.

The K-means clustering algorithm was applied to segment the dataset into distinct groups based on their characteristics. For this, the following steps were followed:

1. Selection and preprocessing the numerical data.
2. Applying the Elbow Method to find the optimal k.
3. Performing K-Means clustering with the identified k.
4. Visualizing the Elbow Method graph.
5. Providing the list of companies and their respective cluster assignments.

The process, as described by Burkardt, begins with selecting initial centroids, either randomly or through advanced strategies. These centroids act as the provisional centers for the clusters. The algorithm then iterates through a two-step process until it achieves stability. Initially, data points are assigned to the closest centroid, typically determined by the Euclidean distance, effectively forming K preliminary clusters. (Burkardt, 2009)

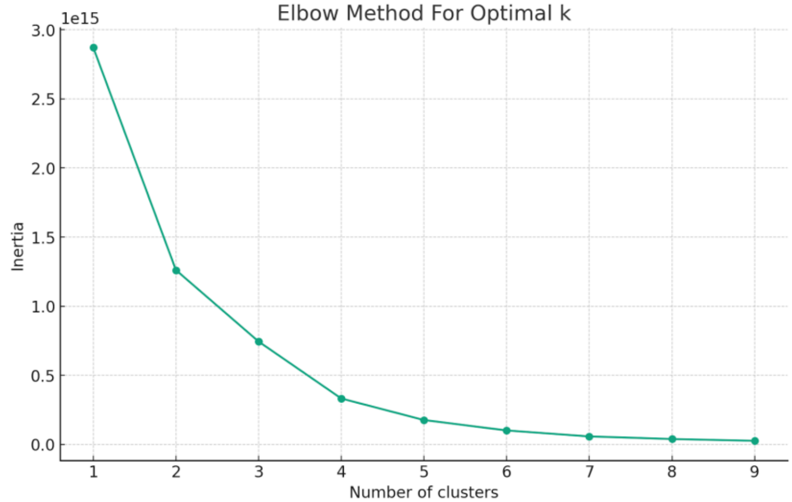

Figure 7. Elbow method for optimal K

Source: Authors' own creation

The Elbow Method graph above illustrates how the inertia (a measure of how well the clustering has been done) decreases as the number of clusters increases. The 'elbow' point, where the rate of decrease sharply changes, suggests the optimal number of clusters for the dataset. Based on this graph, the optimal number of clusters appears to be around 3 or 4, as beyond these points, the inertia decrease becomes more gradual.

Figure 8. Clusters Visualization

Source: Authors' own creation.

The clusters visualization above represents the data points reduced to two dimensions for clarity, with each point colored according to its cluster assignment. This visualization allows us to observe the distribution and separation of clusters within the dataset, based on the K-means clustering with 3 clusters as chosen for demonstration. Each cluster groups data points that are similar to each other, providing insights into the inherent groupings based on the dataset's features.

The application of K-means clustering to the dataset has yielded distinct groupings that reflect variations in company performance, growth trajectories and operational scale. Each cluster encapsulates a subset of companies sharing similar financial and employment characteristics over the analyzed period, providing insights into their market behavior and strategic positioning.

Cluster 0 is indicative of entities exhibiting moderate to high financial growth, with turnovers and net profits showing upward trends. This cluster represents organizations that have successfully navigated market challenges, evidenced by their steady financial performance and workforce expansion. This cluster is characterized by companies undergoing rapid growth, as evidenced by the significant increase in turnover from 2020 to 2022. Despite this growth, these companies report substantial negative net profits, suggesting heavy investments in expansion or development that have yet to yield positive short-term returns. The sharp increase in employee numbers aligns with this growth trajectory, indicating aggressive expansion efforts. The average year of inception points to newer entities, likely startups, in the midst of aggressive growth phases, potentially prioritizing market capture over immediate profitability.

Cluster 1 comprises companies characterized by variability in their financial outcomes, possibly reflecting the volatile nature of startups or firms in early developmental stages. This cluster represents entities experiencing consistent growth, possibly at a more measured pace compared to Cluster 0. The relatively stable and modest increase in employee numbers further suggests these companies are gradually scaling their operations. The average year of inception for these companies indicates they are relatively young but established, likely in a phase of solidifying their market presence and operational efficiencies.

Cluster 2 includes the dataset's most financially successful companies, with the highest turnovers and net profits among the clusters. This financial success is paired with substantial growth in turnover and profit, underlying effective scaling and strong market presence. The relatively stable and high number of employees reflects the large-scale operations typical of well-established firms, corroborated by the oldest average year of inception. This cluster likely represents industry leaders and large enterprises that have developed enduring business models and strategic market approaches.

From figure 8 it results that the majority of analyzed companies pertain to the Cluster 0 category (being 48), followed by five in the Cluster 1 category, only one being in the Cluster 2 category.

Cluster 0 likely encapsulates high-growth, investment-heavy startups facing short-term financial challenges; Cluster 1 includes steadily growing, moderately successful companies; and Cluster 2 represents large, well-established entities with significant financial success and market leadership.

Conclusion

At the core of Industry 4.0 lies a revolution driven by Machine Learning, reshaping the concept of micro and small businesses in Romania. This exploration has revealed how ML serves as a catalyst for operational efficiency and innovation, playing a pivotal role in driving operational efficiency, enhancing predictive analytics and fostering innovation, thus conferring MSEs competitive advantage in the technological landscape. Furthermore, it highlights the critical role of ML in enabling MSEs to navigate the complexities of digital transformation.

Through the application of K-means clustering, the research shows distinct operational and financial profiles across the analyzed enterprises. The findings delimitate three principal clusters, each reflecting different scales of growth and investment strategies, thereby revealing insights into the strategic positioning and market behavior of Romanian MSEs within the digital economy.

In summary, this paper explores the relationship between technology and business, emphasizing the crucial role of digital innovation in shaping modern organizations. With a focus on Machine Learning as a transformative tool, the paper provides a comprehensive overview of its capabilities, types, solutions, and architecture. The initial part underscores the strategic significance of incorporating digital innovation throughout businesses, while the second part delves into practical insights from Romanian micro and small enterprises that have implemented Machine Learning, revealing tangible benefits such as task automation, improved planning, increased profitability and pattern discovery from diverse data models. Overall, the article not only offers theoretical insights into Machine Learning but also provides practical guidance for businesses looking to leverage its potential in navigating the digital landscape. Future research in the area of Machine Learning within Industry 4.0, with a focus on micro and small enterprises in Romania could explore deeper the long-term implications of ML technologies on business sustainability and growth. Possible areas of research could include investigating the barriers of ML adoption and developing strategies for overcoming these challenges.

References

Ahsan, S. (2023, October 11). Getting Started with Deep Learning- A Gentle Introduction. Medium. https://medium.com/@l228122/getting-started-with-deep-learning-a-gentle-introduction-ad5a4a86dc73

Ankitha, S., & Prakash, D. H.N. (2022). System for Recommending Drugs Based on Machine Learning Sentiment Analysis of Drug Reviews. International Research Journal of Engineering and Technology (IRJET), 09(07), 2346-2349. https://www.irjet.net/archives/V9/i7/IRJET-V9I7430.pdf

Bigliardi, B., Colacino, P., & Dormio, A. I. (2011). Innovative characteristics of small and medium enterprises. Journal of technology management & innovation, 6(2), 83-93.

Bouvier, A. (2021). Recommender Systems and Applications in Banking. Medium. https://medium.com/genifyai/recommender-systems-and-applications-in-banking-f0cef8f87fa6

Brown, S. (2021, April 21). Machine learning, explained. MIT Sloan. Retrieved January 27, as-made-to-matter/machine-learning-explained2024, from https://mitsloan.mit.edu/ide

Burkardt, J. (2009). K-means clustering. Virginia Tech, Advanced Research Computing, Interdisciplinary Center for Applied Mathematics.

Kanade, V. (2022, August 30). What is machine learning? Understanding types & applications. Spiceworks. Retrieved January 27, 2024, from https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-ml/

Davis, S. (2017, March 13). What is Net Profit and How to Calculate It. Glew.io. https://www.glew.io/guides/net-profit-calculate

Delua, J. (2021, March 12). Supervised vs. Unsupervised Learning: What’s the Difference? IBM. https://www.ibm.com/blog/supervised-vs-unsupervised-learning/

Domenii CAEN - Lista completa coduri CAEN cu echivalenta NACE si SIC. (n.d.). Client Solutions Group. https://www.clientsolutions.ro/lista-domenii-caen/;

Fedynyshyn, R. (2023, October 16). Price optimization with Machine Learning: Executive's guide. N-iX. https://www.n-ix.com/price-optimization-machine-learning/

Kepuska, V., & Bohouta, G. (2018). Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home). In 2018 IEEE 8th annual computing and communication workshop and conference (CCWC) (pp. 99-103). IEEE.

Lavanya, G., & Praveen, K. (2023). DRUG RECOMMENDER SYSTEM USING MACHINE LEARNING FOR SENTIMENT ANALYSIS. International Research Journal of Modernization in Engineering Technology and Science, 05(07), 2211-2214. https://www.irjmets.com/uploadedfiles/paper/issue_7_july_2023/43424/final/fin_irjmets1690118387.pdf

Lenglet, R. (2023, July 6). Understanding Turnover: Definition and Importance. Agicap. https://agicap.com/en/glossary/turnover-definition/

Microîntreprinderile și întreprinderile mici și mijlocii: definiția și sfera de cuprindere. (2016, January 11). Publications.Europa. https://publications.europa.eu/resource/cellar/1bd0c013-0ba3-4549-b879-0ed797389fa1.0014.02/DOC_2

Shrivastava, A., Krishna, K. M., Rinawa, M. L., Soni, M., Ramkumar, G., & Jaiswal, S. (2023). Inclusion of IoT, ML, and blockchain technologies in next generation industry 4.0 environment. Materials Today: Proceedings, 80, 3471-3475.

Steck, H., Baltrunas, L., Liang, D., Raimond, Y., & Basilico, J. (2021). Deep Learning for Recommender Systems: A Netflix Case Study. AI Magazine, 42(3), 7-18. https://doi.org/10.1609/aimag.v42i3.18140

Suleiman, Z., Shaikholla, S., Dikhanbayeva, D., Shehab, E., & Turkyilmaz, A. (2022). Industry 4.0: Clustering of concepts and characteristics. Cogent Engineering, 9(1), 26. https://www.tandfonline.com/doi/full/10.1080/23311916.2022.2034264

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018, April). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In IOP conference series: materials science and engineering (Vol. 336, p. 012017). IOP Publishing.

The role of AI in content sentiment analysis for customer feedback. (2023, November 6). AIContentfy. Retrieved January 27, 2024, from https://aicontentfy.com/en/blog/role-of-ai-in-content-sentiment-analysis-for-customer-feedback

What is Crunchbase and how does it work? (2022, December 4). Emerging talks. https://emergingtalks.com/what-is-crunchbase-and-how-does-it-work/

What is Data Labeling? (n.d.). IBM. https://www.ibm.com/topics/data-labeling

Zhang, Z., Cheng, Y., & Liu, N. C. (2014). Comparison of the effect of mean-based method and z-score for field normalization of citations at the level of Web of Science subject categories. Scientometrics, 101, 1679-1693.