ИЗПОЛЗВАНЕ НА МАШИННО ОБУЧЕНИЕ ЗА АВТОМАТИЗАЦИЯ ПРИ АНАЛИЗ НА НЕХАРМОНИЗИРАНИ ФИРМЕНИ ДАННИ

Using Machine Learning to Automate the Analysis of Unharmonized Company Data

Теодор Тодоров[1],
e-mail: ttodorov131@unwe.bg [1]

Абстракт
Значително предизвикателство пред малкия и среден бизнес при извличането на знание се явява осигуряването на качествени хармонизирани данни. В малките и средните предприятия информацията често се съхранява във фрагментирани и слабо структурирани файлове, което прави изграждането на надеждни аналитични и прогностични системи сложно и времеемко. Настоящият доклад разглежда възможностите за използване на методи на машинното обучение при работа с нехармонизирани фирмени данни.

Abstract
A significant challenge for small and medium-sized enterprises (SMEs) in the process of knowledge extraction lies in ensuring the availability of high-quality, harmonized data. In SMEs, information is often stored in fragmented and poorly structured files, which makes the development of reliable analytical and predictive systems both complex and time-consuming. This paper explores the potential applications of machine learning methods for working with non-harmonized corporate data.

Ключови думи: извличане на данни, машинно обучение, хармонизация

JEL: C810

Увод

Дори и през 2025 година, малките и средни предприятия продължават да се сблъскват със съществени предизвикателства при анализа на ключова за техния бизнес информация. Непрекъснатото развитие на ERP и BI системите в наши дни води до наличието на все повече данни, които трябва да се превърнат в знание. На пръв поглед, компаниите могат да се възползват от огромен набор информация, чрез която да оптимизират своите процеси, но практиката показва задълбочаване на фрагментацията ѝ от различните платформи. Наличието на разнообразни пазарни решения с ниска месечна цена не разрешава този казус – на разположение са различни - и ефективни сами по себе си – софтуерни решения за счетоводство, продажби, CRM, логистика, управление на финансите. Въпреки всички предимства и по-ниската относителна цена, те често функционират като изолирани острови от данни, без стандартизирани интерфейси и без възможности за автоматизиран обмен на информация помежду си. В резултат на това, бизнес анализът започва с времеемко преобразуване на множество несвързани файлове с разнообразна структура и формати, което изисква значителен човешки ресурс и може да представлява значително предизвикателство пред малки и средни компании.

[1] Докторант, катедра „Информационни технологии и комуникации“, факултет „Приложна информатика и статистика“, УНСС, гр. София, ORCID 0009-0002-5119-8873

Допълнително предизвикателство се явява периодичността в необходимостта от анализи – с приключването на отчетен период или вътрешен организационен етап, компанията е принудена да възпроизвежда целия процес по интеграция и анализ на данните от различните платформи от самото начало. Това води до по-голямо забавяне и – в крайна сметка – игнориране на част от важната информация, която би могла да се превърне в допълнително знание. Разбира се, през последните години платформи като Microsoft Power BI, Tableau, Google Data Studio и други, които помагат за улесняване достъпа до визуализации, като позволяват на потребителите да обединяват данни от различни източници и да изграждат динамични табла за управление. Въпреки това, ефективността на тези решения зависи в голяма степен от предварителното качество и съгласуваност на данните, както и от ресурсите, които компанията може да отдели, за тяхното хармонизиране. Процесът на свързване, настройка на връзки, създаване на дименсии и осигуряване на актуални данни обикновено изисква значителен ръчен труд и техническа експертиза. По тази причина ще бъдат разгледани възможностите за различни подходи за автоматизирано извличане на знание от нехармонизирани фирмени данни, при които машинното обучение може да заеме ключова роля в ETL процеса.

Ключови положения при анализ на данни в малки и средни предприятия

Подготовка на данни. Хармонизирани и нехармонизирани данни

В свое определение, изданието Gartner (2023) поставя акцент върху непрекъснатата повтаряемост на процеса по предварителната подготовка на данни. Това е сложна и многопластова дейност, която изисква значителна човешка намеса – и специалисти по данни, и бизнес анализатори, които да дадат възможност за извличане на знания за бизнес нуждите на организацията. Дори при използването на модерни инструменти за автоматизация на визуализациите, в основата на анализа остава ръчната работа по почистване и съгласуване на данните. Именно това обуславя и нуждата от нови подходи, а една възможност е машинното обучение да поеме част от тези задачи чрез автоматично откриване на зависимости, аномалии и липси.

В съвременната литература, терминът хармонизирани данни, harmonized data, обозначава съвкупност от различни по смисъл данни, които са съгласувани в „технически“ смисъл – имат еднакъв формат, структура, семантика. В този ред на мисли, различните източници на информация трябва да имат сходна логика на описание, позволяваща анализ и интерпретация (Inmon, 2005; Kimball and Ross, 2013). Хармонизацията има за цел да осигури взаимна съвместимост между системи, така че една и съща бизнес променлива – например приход – да бъде разбирана по еднакъв начин в различния бизнес контекст. В своя разработка от 2021 година, Николай Драгомиров и Любен Боянов (Dragomirov, Boyanov, 2021) изследват важни предизвикателства при анализа на данни от различни източници в сферата на логистиката, като подчертават възможностите за огромно количество потенциално знание. За да бъде достигнато то, обаче, стои необходимостта от уеднаквяване (хармонизация), което се извършва в по-голяма платформа за анализ на големи данни. Чрез хармонизация на данните, може да бъде постигната и „крайната цел“ в сферата - адекватен анализ. Разбира се, от другата страна, или по-точно – в началото на всяко намерение за анализ, стоят т.нар. „нехармонизирани данни“. Те са дефинирани като неструктурирани, несъвпадащи или частично-съвпадащи единици от информация, които се явяват основно предизвикателство пред автоматизираната обработка (Goodfellow, Bengio and Courville, 2016). Различия между данните често възникват още в създаването на отделните „късчета“ информация в ERP системите, тъй като различни

части на платформите могат да използват различни условности или формати за вход-изход. Липсата на автоматична проверка и централизирано управление на първичните данни може да доведе до т.нар. разминаване на значенията, semantic drift – в този случай, освен че данните стават трудни за интерпретация, те могат да доведат до възникването на неточни изводи (Ivanova, Iliev, Stoyanov, 2020). В свой труд за интернет на нещата (IOT), четирима български автори отбелязват съществената необходимост от хармонизация на данните като предпоставка за изграждането на интегрирани модели за анализ (Zlatev, Georgieva, Todorov, Stoykova, 2022). В същата разработка се отбелязва и че липсата на хармонизирани данни в повечето български компании може да бъде пряка последица от историческата зависимост към универсални инструменти (като Excel), което в контекста на настоящия доклад е съществена пречка пред автоматизацията анализа на ключови показатели

В следващите части на доклада, ще бъде обърнато внимание именно на някои автоматизирани подходи за хармонизация, при които алгоритми за машинно обучение могат да изпълняват задачи като:

- Автоматизирано сравнение имената на колони с различни наименования, т. нар. schema matching;
- Разпознаване и преобразуване на различни мерни единици, unit conversion;
- „Приписване на значения“ към данни с липсващи стойности, data imputation;
- Откриване и корекция на аномалии, anomaly correction.

Тези подходи биха могли да съкратят времето и разходите за хармонизация на данни и биха били особено полезни за малки и средни предприятия, като прехвърлят значителна част от „мисловната“ дейност от човека към компютърната система (Agrawal, 2023).

Съществуващи решения и ограничения.

Клиентски BI и ETL системи с модули с изкуствен интелект

Голяма част от малките и средни компании вече са добре запознати с текущите решения за визуализация на данни на пазара. Освен тази си основна функция, продукти като Power BI и Data Studio предлагат помощ при събиране и трансформация на данни. Добавяйки факта, че става дума за сравнително евтини решения с месечен абонамент (Software as a service, SaaS модел), малкият бизнес често е „изкушен“ от това да започне с някои от изброените решения. Част от възможностите на подобни BI системи са именно връзките с различни източници на данни – Excel файлове, извличане на данни чрез API и други. Основното предизвикателство, обаче, отново се корени във въпроса какво да направи предприятието с данните, преди да ги „подаде“ към BI платформата. Редица автори посочват, че тези платформи не решават основния проблем с качеството и съгласуваността на данните. В свое практическо изследване от 2021, няколко индийски учени изказват тезата, че въпреки изключителната си визуална мощ, BI инструментите остават зависими от предварително обработени и надеждни данни, без които техните изводи губят стойност (Vasudevan, 2021). Ако се опитаме да оборим това твърдение с наличието на все по-напреднали допълнения (add-ons) с изкуствен интелект в решенията на водещите софтуерни компании, ще установим, че дори и днес изкуственият интелект все още има нужда от редица насоки и потвърждения от потребителя, преди да пристъпи към конкретно действие. Добавяме и възможността от

допускане на голям брой еднотипни пропуски и автоматизацията на този процес се оказва подпомогната, но не и постигната, и продължава да бъде налице необходимост от тясно проследяване на всеки аспект от хармонизацията от страна на бизнес организацията. Казано накратко, макар решенията за бизнес анализи да стават все по-мощни и многопластови, те продължават да са последна инстанция от анализа на данни и са най-полезни при визуализацията на вече обработени данни. Дори и при наличие на инструменти с изкуствен интелект, се изисква от потребителя предварително да дефинира цялостната логика зад данните, която той трябва да разбира.

Low-code платформи за интегриране на данни

Агресивното развитие на облачните технологии спомогна и за утвърждаването на корпоративни платформи от тип Low-code. Това са софтуерни решения за ускорено разработване и поддръжка на приложения, използващи инструменти за разработка, базирани на модели; генеративен изкуствен интелект и предварително изградени каталози на компоненти за целия технологичен стек на приложението (Gartner, 2025). Примери за подобни продукти са Talend, Apache NiFi, Airbyte, RapidMiner, като те предоставят възможности за цялостно моделиране на ETL процеси и често се използват в научни и големи корпоративни среди за бързо моделиране на анализи. Low-code и open-source платформите като KNIME, RapidMiner и Talend имат за цел да улеснят потребителите чрез визуални среди за моделиране на процеси и готови конектори. Въпреки това, тяхното ефективно използване изисква концептуално разбиране на моделите на данните и основни умения за интеграция, което често надхвърля капацитета на малките предприятия (Gartner, 2024; Deloitte, 2022). В този смисъл, макар да елиминират нуждата от писане на код, те не елиминират нуждата от експертиза в анализа на данни и логиката на ETL процесите.

Умни ETL решения. AI-Enhanced ETL.

С нарастването на обемите от данни и необходимостта от бърз анализ, през последните години възниква ново поколение технологии, определяни като AI-Enhanced ETL (разширени чрез изкуствен интелект процеси по трансформация на данни, Intelligent Data Preparation Tools). Те намаляват времето за хомогенизиране и интеграция на фирмените данни, като позволяват на бизнес потребителите да използват едновременно различни източници. Освен това, те дават шанс на компаниите да идентифицират аномалии, да изследват и очертават модели, да подобрят качеството на данните от своите открития (Gartner, 2025). Такива инструменти използват изкуствен интелект и машинно обучение, за да автоматизират традиционните фази на извличане (Extract), трансформация (Transform) и зареждане (Load) на данни.

Класическият ETL процес работи по предварително зададени правила – всеки източник на данни трябва да бъде описан, всички полета – съпоставени, трансформациите – дефинирани.

AI-Enhanced ETL добавя **слой от интелигентност**, при който платформата:

- Автоматично **определя и категоризира** данните – открива типове полета, формати и зависимости;

- **Открива аномалии** и липси чрез алгоритми за машинно обучение (напр. чрез Isolation Forest, AutoEncoder);
- Извършва **автоматична импутация** на липсващи стойности на базата на вероятностни зависимости;
- **Прилага трансформации**, използвайки модели, обучени върху предишни операции;
- **Оптимизира последователността** на задачите.

По този начин системата не просто изпълнява инструкции, а се самообучава от предишни трансформации и грешки, намалявайки нуждата от човешка намеса при повторно извличане на данни. AWS Glue DataBrew, Google Cloud Dataprep (Trifacta), Snowflake Cortex и Databricks AutoML, например, прилагат елементи на ML за автоматично профилиране, откриване на аномалии и генериране на препоръки за трансформации. Според Gartner (2024), *„умните“ ETL инструменти се намират във фаза на преход от асистираща към автономна автоматизация*, като повечето решения все още предлагат препоръки, но не ги прилагат самостоятелно. С други думи, AI-Enhanced ETL все още не е напълно самообучаваща се система, но бележи първата стъпка към създаването на процеси, които автоматично се адаптират при промени в структурата или съдържанието на данните.

Въпреки напредъка, прилагането на изкуствен интелект в процесите по трансформация и зареждане на данни, среща няколко предизвикателства:

- Необходимост от вече налични големи обеми исторически данни за обучение на моделите;
- Възможни затруднения при интерпретацията на автоматично извършените трансформации (т.нар. explainability проблем – не е ясно защо „платформата“ взима конкретно решение);
- Повишени изисквания към изчислителните ресурси и сигурността (макар този казус да се разрешава сравнително лесно с използването на облачни ресурси)

В контекста на малките и средни предприятия, тези ограничения често правят внедряването им неефективно от финансова гледна точка. Въпреки това, тенденцията е ясно очертана – бъдещето на ETL процесите е в самообучаващите се модели, които ще позволят анализ върху нехармонизирани данни с минимална намеса на човека.

Направеният обзор показва, че независимо от напредъка в областта на бизнес анализа и автоматизацията на данните, съществуващите решения остават зависими от предварителна подготовка и ръчна намеса. Дори и най-модерните инструменти, които използват машинно обучение, все още не притежават пълна автономност и не могат самостоятелно да обработват нехармонизирани фирмени данни в тяхната първоначална форма. Това създава реална нужда от методология, която да обедини предимствата на ML-базираните алгоритми с практичността на BI инструментите, като по този начин осигури еднократно изграждане и последващо самообновяване на анализа при постъпване на нова информация. Следователно, в следващата глава се предлага рамка за автоматизация на анализа на нехармонизирани фирмени данни, основана на интеграция между алгоритми за машинно обучение и процесите на класическия ETL цикъл.

Анализ на възможностите за автоматизация при хармонизацията на данни чрез машинно обучение.

Още през 1959, Артър Лий Самюъл дефинира машинното обучение като наука, която изучава способността компютрите да се обучават за изпълнение на задача, без да бъдат конкретно програмирани за извършването на същата. След това, през 1986, Том Мичъл развива съвременната дефиниция за машинно обучение, като я представя като зависимост между ефективността за изпълнение на определена задача, и увеличаването на опита, натрупан от изпълнението на същата. Съвременните алгоритми могат да открият закономерности, групи и отклонения в данни с неоднородна структура и да подпомагат ETL процеса. Така например, модели като *k-nearest neighbors (KNN)* и *Multiple Imputation by Chained Equations (MICE)* се използват за импутация на липсващи стойности (Azur, 2011), докато *Isolation Forest* и *AutoEncoder Neural Networks* са ефективни за откриване на аномалии в таблични данни (Sakurada & Yairi, 2014). Тези методи позволяват на системата автоматично да допълва, коригира и валидира данните, намалявайки времето, необходимо за тяхното ръчно преглеждане. В контекста на малкия и среден бизнес прилагането на машинно обучение има стратегическо значение, защото премахва нуждата от дълбока техническа експертиза, характерна за големите организации. Както посочват Gartner (2024), нарастването на обемите от данни и разпространението на визуални инструменти за машинно обучение създават възможности за „демократизация“ на анализа – т.е. достъп до автоматизирани функции без специализирани познания по програмиране. Такива платформи могат да извършват *schema matching* (съпоставяне на несъответстващи колони), *unit normalization* (уеднаквяване на мерни единици и формати) и *semantic labeling* (автоматично приписване на значения към колони) въз основа на вече обучени модели (Wang, 2022). Тези функционалности превръщат машинното обучение в „интелигентен посредник“ между различните източници на информация и позволяват по-бързо достигане до консистентни, готови за анализ данни. Освен в ETL процеса, ML алгоритмите могат да бъдат използвани и за прогнозен анализ на ключови бизнес индикатори. Модели като *XGBoost* позволяват предсказване на тенденции на базата на исторически данни, като се адаптират автоматично към сезонни колебания и липсващи наблюдения – нещо, което традиционните статистически методи трудно постигат. Това е особено полезно за компании, които не разполагат със собствени отдели за анализ, но се нуждаят от бърза и достъпна прогноза за бъдещи продажби, разходи или пазарни рискове.

Процесът на хармонизация чрез машинно обучение обединява няколко взаимосвързани стъпки, които могат да се прилагат последователно или паралелно. Първата стъпка е *профилиране на данните*, при което алгоритми за клъстериране и класификация (напр. *k-means* или *decision trees*) автоматично откриват сходства между колони с различни наименования и определят каква е тяхната семантика — например „Client ID“, „Customer_Number“ и „Клиент“ се разпознават като идентификатори на едно и също понятие. След това се прилагат модели за *нормализиране на стойности*, които уеднаквяват формати и единици, като се обучават върху примери от предишни трансформации (Abdelaal, 2023). В третата стъпка се използват методи за *импутация* и *откриване на аномалии*, при които машинното обучение идентифицира липсващи или нетипични стойности и предлага вероятни корекции на базата на сходни наблюдения (Azur, 2011). Накрая, чрез *self-supervised* подходи, системата може да изгражда собствени представяния на данните (*embeddings*), които улесняват последващи съпоставяния между различни източници, дори когато имената и формите на колоните са напълно различни. По този начин хармонизацията престава да бъде изцяло ръчен

процес и се превръща в непрекъснат цикъл на самоусъвършенстване, при който моделите постепенно „научават“ структурата на корпоративните данни и я прилагат автоматично при нови източници.

Заключение

В ерата на нарастващи обеми от информация малките и средните предприятия са изправени пред парадокс – разполагат с повече данни от всякога, но трудно ги превръщат в знание. Липсата на хармонизирани източници, разнообразието от формати и ограничените ресурси за поддръжка на сложни ИТ решения правят традиционните подходи за анализ неефективни. В този контекст машинното обучение предлага все по-реалистична и гъвкава алтернатива. В природата на ML е да открива закономерности, да попълва липси и да уеднаквява разнородни структури, като така значителна част от ръчната работа по подготовка и интерпретация на данните може да отпадне. Настоящият доклад разгледа концепцията за използване на алгоритми за машинно обучение при анализа на нехармонизирани фирмени данни, като очерта възможностите за автоматизация на ключови етапи от процеса – профилиране, нормализиране, импутация и откриване на аномалии. Представеният теоретичен модел демонстрира, че интеграцията между съществуващите BI инструменти и ML-базирани алгоритми може да изгради единен и самообновяващ се аналитичен цикъл, който съкращава времето за анализ и повишава надеждността на резултатите. Така малките и средните предприятия могат да използват предимствата на изкуствения интелект без необходимост от мащабна ИТ инфраструктура или сериозни програмни умения.

В по-широк план машинното обучение се явява не просто технологична иновация, а методологичен мост между данните и бизнес знанието. То позволява на организациите да преминат от еднократен, ръчно извършван анализ към непрекъснат процес на самоусъвършенстване, при който всяка нова информация автоматично подобрява точността на следващата. Това поставя основата за следващи изследвания, насочени към практическа реализация на предложената рамка и към разработване на достъпни, адаптивни решения за анализ и прогнозиране на данни в реалната бизнес среда на малките и средни предприятия.

Източници

1. Agrawal, R., Majumdar, A. & Kumar, A. et al. (2023). *Integration of Artificial Intelligence in Sustainable Manufacturing: Current Status and Future Opportunities. Operations Management Research*
2. Azur, M.J., Stuart, E.A., Frangakis, C. & Leaf, P.J. (2011). *Multiple Imputation by Chained Equations: What Is It and How Does It Work? International Journal of Methods in Psychiatric Research*. Available at: https://pubmed.ncbi.nlm.nih.gov/21499542/ (Accessed: October 2025).
3. Boyanov, L. & Dragomirov, N. (2021). *Supply Chain Management and Logistics Big Data Challenges in Bulgaria. Academia.edu.* Available at: https://www.academia.edu/68932279/Supply_Chain_Management_and_Logistics_Big_Data_Challenges_in_Bulgaria (Accessed: October 2025).
4. Chen, T. & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

5. Gartner (2023). *Market Guide* for *Data Preparation Tools.* Gartner Inc. Available at: https://www.gartner.com/en/documents/3987296 (Accessed: October 2025).
6. Gartner (2024). *Enterprise Low-Code Application Platforms. Gartner Peer Insights.* Available at: https://www.gartner.com/reviews/market/enterprise-low-code-application-platform (Accessed: October 2025).
7. Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep Learning.* MIT Press. Available at: https://www.deeplearningbook.org (Accessed: October 2025).
8. Inmon, W.H. (2005). *Building the Data Warehouse* (4th ed.). Wiley.
9. Ivanova, E., Iliev, T. & Stoyanov, I. (2020). *Modelling and Simulation of Big Data Networks. IEEE Xplore*
10. Kimball, R. & Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* (3rd ed.). Wiley.
11. Liu, F.T., Ting, K.M. & Zhou, Z.-H. (2012). *Isolation Forest. ACM Transactions on Knowledge Discovery from Data.*
12. Mittal, V. (2025). *Leveraging Artificial Intelligence to Optimize ETL Pipelines: Enhancing Efficiency, Accuracy and Scalability. ResearchGate preprint.* Available at: https://www.researchgate.net/publication/381925765_Leveraging_AI_to_Optimize_ETL_Pipelines (Accessed: October 2025).
13. Mitchell, T.M. (1997). *Machine Learning.* McGraw-Hill Education.
14. Sakurada, M. & Yairi, T. (2014). *Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction. Proceedings of the 2nd Workshop on Machine Learning for Sensory Data Analysis (MLSDA 2014).*
15. Schrage, M. (2024). *The Future of Strategic Measurement: Enhancing KPIs With AI. MIT Sloan Management Review.* Available at: https://sloanreview.mit.edu/projects/the-future-of-strategic-measurement-enhancing-kpis-with-ai (Accessed: October 2025).
16. Soon, E., Vasudevan, A., Alshamayleh, A. & Mohammad, S. (2025). *Redefining Realities: AI's Cost-Efficient Revolution in Indian Cinema's VFX Landscape. Natural Sciences Publishing.* Available at: www.naturalspublishing.com/files/published/519fn13620io92.pdf (Accessed: October 2025).
17. Taylor, S.J. & Letham, B. (2018). *Forecasting at Scale. The American Statistician.*
18. Wang, Z., Li, H. & Wang, C. (2022). *Sudowoodo: Contrastive Self-Supervised Learning for Multi-purpose Data Integration and Preparation.* Available at: https://arxiv.org/abs/2207.04122 (Accessed: October 2025).
19. Zlatev, Z., Georgieva, Ts., Todorov, A. & Stoykova, V. (2022). *Energy Efficiency of IoT Networks for Environmental Parameters of Bulgarian Cities.*