Intellectual Technologies in Digital Transformation

P A Sakhnyuk¹, T I Sakhnyuk²

¹National Research University Higher School of Economics, 20, Myasnitskaya str., Moscow, 101000
²Moscow City University, 4, 2-nd Selskohozoyastvenny street, Moscow, 129226

E-mail pav-sahnyuk@yandex.ru

Abstract. Artificial intelligence and machine learning helps to improve the quality of customer service and change the methods of companies’ activities. For this reason, enterprises should consider integrating these technologies into digital transformation plans to remain competitive. Low-code machine learning platforms allow companies and business professionals with minimal coding experience to create applications and fill in the gaps of the personnel in their organization. Automated machine leaning (AutoML) technology represents the next step in the evolution of machine learning, providing non-technical companies with the ability to create machine learning applications quickly and cheaply.

Keywords: Digital transformation, data science and machine learning platforms, automated machine learning, artificial intelligence.

1. Introduction
Information is a key resource in the contemporary world. Humanity generates huge amounts of digital data, which not only occupy a place in cloud storage and data lakes but also help companies have done business. Collection, storage, structuring and analyzing data are necessary to take full advantage of the available information. The variety, complexity and speed of this data as well as its many sources changes the methods of companies’ activities. Perhaps, the biggest factor in data growth is advances in technology that have made data collection and data exchange easier and more accessible. Modern users have also become more technically competent and use digital technology to create and share more data and applications [1-2].

Digital transformation is a strategic transformation aimed at a business can operate in a rapidly changing environment, identify key goals and develop the trajectories necessary to achieve these goals. Digital transformation allows integrate digital technology in all business areas, so business processes can change for the better. Digital transformation achieves its goals due to advanced technologies such as big data (BD) or artificial intelligence (AI). These technologies are aimed at processing information flows, on the basis of which it is possible to make decisions, adapt offers to specific clients and predict their behavior. AI and machine learning (ML) process and analyze data quickly and accurately. These technologies are a key part of revealing the value of company and customer data, allowing make decisions faster and clearly.

2. Informal statement of the problem
AI is one of the main areas of activity and the main engine of digital transformation. It has become an excellent tool for companies undergoing digital transformation. Data grows exponentially, and with it
grows the potential for extracting valuable business information from them, understanding problems and automating routine operations using ML.

The development of AI is pushing digital enterprises to transform into intelligent enterprises. This is mainly due to advances in ML combined with technology based on data science. AI technologies are already revolutionizing not only how we perceive and conduct business, but also the business environment and its overall IT landscape. Many companies invest in the development of AI technologies, but they are hampered by the lack of skills in the practical application of the developed models.

Unfortunately, ML models are often expensive to develop and require a lot of data to build and train algorithms. Traditional enterprise-level AI/ML implementation methods are complex and typically require expensive, highly trained Data Science professionals. The widening gap in the qualifications of business users and Data Science professionals is also creating problems for organizations. Smart technology projects can include months of data preparation, followed by the design, development, testing, tuning, and deployment of ML models.

It is also important that many developed solutions based on AI/ML are difficult to trust. They consist either of complex non-linear models and ensembles of models that are difficult or almost impossible to interpret by people, or of pre-built models that do not give an idea of how the predictions are made. When using the “black box” model, the input data are known and receive the output data, but it is not known what happens between them. This model may not be acceptable because it is important that the business trust your results. Understanding the drivers of certain behaviors is what allows you to make the right decisions and have a real impact on the business. That is why it is important to work in cross-functional teams with industry experts from the business environment. As AI and ML become an integral part of any part of the business, it is important that all employees have access to the tools and solutions that ML offers. It is clear that experts in the field of business, who have knowledge in business to identify the main problems and opportunities, need access to these tools. However, they do not have technical expertise. In other words, they cannot write code.

Thus, today there are three big problems in the application of intelligent technologies in digital transformation: lack of data and analytics experts; a long and expensive period of development and implementation of machine learning models; trust in the results obtained with their help. All of these together prevent organizations from implementing AI/ML for digital transformation at the enterprise level.

3. The ability to use the data science and ML platforms and automation-focused ML.

Nonetheless, modern solutions use advanced ML capabilities even for companies with minimal experience in data processing. The accelerated implementation of AI and ML, combined with the availability of open-source software, allows you to create and train far more models than ever. Open-source software allows IT departments to access infrastructure, data sets, workflows and trained models (transfer training). This approach reduces the overall cost of implementing intelligent technologies by providing developers with tools for developing models and using the experience of other Data Science specialists. This is why companies are interested in Citizen Data Scientists. These new data processing specialists are closing the skills gap on the part of the business using automated and other powerful data processing technologies [3-5].

The term "open source software" refers to a tool with source code available over the Internet for free. The proprietary (closed) program code is private and distributed through licensing rights. For a business that is just starting its ML initiative, using the open-source tools Knime (Figure 1), H2O.ai, RapidMiner can be a great way to practice data science for free before choosing enterprise-level tools such as machine service Amazon machine learning et al.
Figure 1. Workflow Knime trains a simple convolutional neural network on the MNIST dataset through Keras

Figure 2. Web interface of Microsoft Azure Machine Learning Studio
The benefits of using open source tools are not limited to their availability. Typically, open-source projects have a huge community of data engineers and scientists seeking to share data sets and pre-trained models.

Data Science and Machine Learning Platforms tools use ML algorithms to scale Data Science workflows. Data Science project teams using platforms that combine data preparation, ML, and model deployment are more productive. Automated platforms scale the capabilities and resources of the data scientists team by providing enhanced functionality for data visualization, feature development, model interpretation, and low-latency deployment.

Among the Data Science and Machine Learning Platforms, platforms with Visual workflow designer, for example, Microsoft Azure Machine Learning Studio (Figure 2), are of great interest. Visual Workflow Designer allows increasing the productivity of the entire data scientists team, from analysts to experts, to speed up and automate the creation of predictive models in the drag and drop visual interface. To fine-tuning the models of the platform, the R & Python scripting nodes written in the Jupyter Notebook, specialized and advanced ML libraries are combined in one workflow, providing a developed solution for the rest of the team.

Low-code is a visual approach to application development. Low-Code Machine Learning does most of the development work using drag-and-drop components and model-driven logic through a graphic user interface. Low-Code ML frees non-technical developers from having to write code while supporting professional developers while abstracting from the tedious infrastructure tasks required when developing applications. Working together Citizen Data Scientists and of data and analytics experts create, train and deploy models in an order of magnitude less time than traditional methods. Low-Code Machine Learning creates a full range of applications for solving problems of classification, regression, clustering, anomaly search and recommendation systems [6-10].

Tools such as Visual workflow designer in Azure Machine Learning and automatic ML capabilities are changing this paradigm. ML used to be a black box or was perceived as unattainable, very complex and unusable by anyone who could not code.

Previously, deploying a trained model in a production environment required the experience of a data specialist, knowledge of coding, model management, container maintenance, and web service testing. Using the constructor, the user can drag and create an experiment and then run it, which means that any business user can deploy a trained model with just a few clicks. This provides the real value of AI / ML projects for business.

4. The ability to use the Automation-Focused Machine Learning
Much of the time in a typical data science project is devoted to building a model or choosing a model. This is a complex task requiring expensive calculations and time consuming data scientists.

Using representative structured data (flat file tables), a machine learning algorithm can build a model of what it expects from the data. Then it compares the results of the model with the reference data, finds the error as the difference between the target and actual values, and iteratively modifies its parameters to minimize the error. When choosing the right settings, taking into account and evaluating all possible features and basic trainees of the ensemble model, an optimal conveyor is obtained that provides the highest possible accuracy [11, 12].

The complex and iterative nature of the feature development makes qualified data scientists relevant, making it difficult to find them. At present, ML as technology has reached a point where the development of features and other important stages of model training can be performed automatically with a high degree of accuracy. This allows organizations to complete this important ML workflow faster by focusing on data scientists on the preparation of the presentation sample to be trained.

If the target or data changes (for example, new attributes are added), the process must be repeated. AutoML can help data scientists save time searching for the best model and its hyperparameters by focusing on a better understanding of the business problem being solved and the quality of the data.

An input for AutoML is data and a task (classification, regression, recommendations, etc.), an output is a model ready for production, capable of qualitatively solving forecast problems with new data. Hyperparameters of a model are parameters whose values are set before the model starts training and does
not change during the training process. Each solution in a data-driven pipeline is a hyperparameter. The task of AutoML is to find such hyperparameters that could give a good result in a reasonable time though:

- to choose a strategy for data preprocessing: how to work with unbalanced data; how to fill in the missing values; remove, replace or save outliers and anomalies; how to encode the values of a categorical variable; how to avoid data leakage from input characteristics to the target variable, etc.
- to generate new feature and choosing the most significant;
- to choose the best model (generalized linear models, XGBoost, Random Forest, deep neural networks, etc.);
- to search for hyperparameters of the selected model (for example, the number of trees and parts of the dataset for models or architecture based on trees, learning speeds and the number of eras for neural networks);
- to create an ensemble of models to increase the accuracy of the forecast, if possible.

![Figure 3. Web interface of H2O FLOW](image)

The only free library for industrial use with automatic machine learning is H2O AutoML. The library provides an easy-to-use interface that automates data preprocessing, training and tuning a large number of candidate models (including models with several stacks for excellent model accuracy and performance). The result of AutoML is the “Leaderboard” of H2O models, which can be easily exported for use in production. AutoML is available in all H2O interfaces: R, Python, Scala, the web interface (Figure 3) and due to the distributed type of the H2O platform, it can be scaled across clusters to process big data. The open-source ML components H2O.ai are an industry standard that integrates many other platforms (for example, Alteryx, Dataiku, Domino, IBM, KNIME, RapidMiner, and TIBCO Software) [13-15]. The library also supports multi-core processor architectures and offers a software layer for significantly accelerating the GPU.

AutoML H2O.ai key features are:

- automatic selection of algorithms and tuning automates the test execution process for several algorithms and hyperparameter configurations. AutoML checks the results for accuracy and confirms that the optimal model and configuration has been selected for use. This significantly saves the time of data
processing and analysis specialists and, more importantly, allows each of them to get the same results as the most experienced specialists;

- automatic selection of predictive features simplifies the creation and selection of features by automatically identifying key predictive features from large data sets;
- model assessment generates a complete set of assessment metrics and related visualizations to measure model characteristics with new data. It allows ranking models over time to ensure optimal behavior of the working version. Assessing a model is beyond the scope of a direct assessment of performance. In order to fully take into account the various effects of errors of the first and second kind (false positive and false negative), the expected basic behavior is taken into account and the cost model is used;
- model explanation: automatically provide explanations of the relative weight and importance of factors affecting the formation of the forecast.

5. Conclusion

Data Science and Machine Learning Platforms and Automated ML simplifies the process of developing the ML model and gives business users, regardless of their experience in the field of data science, the opportunity to define an end-to-end machine learning pipeline for solving Data Mining problems. Automated ML makes it easy for Citizen Data Scientist to create machine learning models without writing code. Automated ML includes the best machine learning practices from leading experts in the field of Data Science, and modern open-source libraries to make machine learning more accessible to the entire organization.

Designing a model using a traditional process is extremely time-consuming, repetitive and tedious. The AutoML application automatically performs model building tasks that typically require qualified Data Scientists. Instead of taking weeks or months, the automated ML system is fast and usually requires business users / analysts to have these days to create hundreds of models, predict and analyze. Automated ML for data analysts enables organizations to do more in less time.

AutoML allows enterprises in industries such as healthcare, FinTech, banking, and others to use advanced machine learning and artificial intelligence technologies that were previously limited to large business companies. By automating most of the tasks of training analytical models, AutoML allows business users and data analysts to easily integrate into business processes and pay more attention to solving complex business problems.

Data Scientists and business users from various industries can use automatic machine learning with a view:

1. to implement ML-based solutions without in-depth programming knowledge.
2. to save time and resources.
3. to use best practices in data science.

By providing accurate forecasts and recommendations, applications based on Low-Code Machine Learning and AutoML can improve customer service, improve product quality and reduce costs. For these reasons, enterprises should consider integrating these technologies into digital transformation plans in order to remain competitive.

6. Acknowledgment

The work is partially supported by the RFBR grant # 18-413-770006.

References

[1] Frolov Y V, Sakhnyuk P A and Sakhnyuk T I 2018 Use of tools of geoanalytics for training of bachelors and masters in the business informatics direction Rudn J. Informatiz. Educ. 15 151–64
[2] Mamaev I, Sakhnyuk T and Sakhnyuk P 2016 Analysis of the main socio-economic indexes of districts in the Stavropol region in 2011-2015 by means of business analytics platforms Sci. J. KubSAU 9 1–14
[3] Siddiqui T and Ahmad A 2018 Data mining tools and techniques for mining software repositories: A systematic review Advances in Intelligent Systems and Computing vol 654 (Springer Verlag) pp 717–26
[4] Kanupriya V, Sahil B, Resham A, Mir S, Ul I, Megha B, Ashok K and Samant P 2019 Latest
Tools for Data Mining and Machine Learning Int. J. Innov. Technol. Explor. Eng. 8 18–23

[5] Alcalá R, Gacto M J and Alcalá-Fdez J 2018 Evolutionary data mining and applications: A revision on the most cited papers from the last 10 years (2007–2017) Wiley Interdiscip. Rev. Data Min. Knowl. Discov.

[6] Jean N, Burke M, Xie M, Davis W M, Lobell D B and Ermon S 2016 Combining satellite imagery and machine learning to predict poverty Science (80-. ).

[7] Froelich W 2017 Towards improving the efficiency of the fuzzy cognitive map classifier Neurocomputing

[8] Mitchell R Gradient Boosting, Decision Trees and XGBoost with CUDA | NVIDIA Developer Blog

[9] Mitchell R and Frank E 2017 Accelerating the XGBoost algorithm using GPU computing PeerJ Comput. Sci.

[10] Shi S, Wang Q, Xu P and Chu X 2017 Benchmarking state-of-the-art deep learning software tools Proceedings - 2016 7th International Conference on Cloud Computing and Big Data, CCBD 2016

[11] Chen T and Guestrin C 2016 XGBoost: A scalable tree boosting system Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining

[12] Nguyen G, Dlugolinsky S, Bobák M, Tran V, López García Á, Heredia I, Malik P and Hluchý L 2019 Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey Artif. Intell. Rev.

[13] Hall P 2018 On the Art and Science of Machine Learning Explanations KDD ’19 XAI Workshop, August 04–08, 2019, Anchorage, AK

[14] Hall P, Gill N and Schmidt N 2019 Proposed Guidelines for the Responsible Use of Explainable Machine Learning 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada

[15] Slack D, Friedler S A, Scheidegger C and Roy C D 2019 Assessing the Local Interpretability of Machine Learning Models arXiv Prepr. arXiv1902.03501