EVALUATION OF THE PERFORMANCE OF AUTOMATED MACHINE LEARNING TOOLS

Desislava Koleva, PhD student[1], Assoc.Prof. Yanka Aleksandrova, PhD[2]

[1] University of Economics, Department of Informatics, Varna, Bulgaria, desi_koleva@ue-varna.bg

[2] University of Economics, Department of Informatics, Varna, Bulgaria, yalexandrova@ue-varna.bg

ABSTRACT

The ubiquitous application of predictive models has created the demand for optimized ways of building, deploying and enhancing machine learning models. Traditionally building and deploying machine learning models require the involvement of highly classified data scientists with good knowledge about machine learning algorithms, specialized programming languages, mathematics, statistics and data engineering.

The relatively new and developing area of automated machine learning (AutoML) has made the whole process of building and deploying an AI model more accessible and automated by providing solutions for automated machine learning pipeline encompassing data collection, preprocessing, feature selection, model building and hyperparameter tuning. Automated machine learning tools allow for business users that are not necessarily machine learning experts to develop and implement high quality predictive models.

The purpose of this research paper is to assess and compare AutoML tools for solving classification problems. One of the leading AutoML tools are chosen like Azure Automated Machine Learning, Amazon Sage Maker Auto Pilot, H2O AutoML, H2O Flow and Altair AI Studio Auto Model. Classification models using AutoML tools are trained on a dataset for customer churn predictions and models are compared based on previously chosen measures. AutoML tools are also assessed on different criteria like ease of use, functionality, user orientation, limitations and generated output format. Results show remarkable predictive performance of the generated classification models in general. The best trained classification models are ensemble models trained in Amazon Sage Maker Auto Pilot and H2O AutoML. Some conclusions and recommendations have been drawn to help data science practitioners with choosing and implementing automated machine learning tools.

KEYWORDS: *AutoML, automated machine learning, Azure, Amazon SageMaker, H2O, Altair*

INTRODUCTION

In the current business environment, machine learning, as a field of artificial intelligence, is the main approach for researching the presence of dependencies between various factors involved in a certain process or occurrence. Recently, the use of automated machine learning (Auto ML) tools has become popular. It enables the automatic solution of real-world problems through machine learning techniques (Consuegra-Ayala, et al., 2022) and solves some problems associated with the conventional use of machine learning (ML) methods including time consuming, resource intensive, manually feature engineering, multiple algorithms testing, optimizing many of model parameters, hiring experienced data scientists. It automates end-to-end machine learning implementation and supports all stages of ML pipeline incorporate data collection, data preprocessing, training, tunning, evaluation, explanation and deployment. Graphic user interface (GUI) is one of the benefits of Auto ML that provides dialog boxes and wizards with subsequent steps during the Auto ML pipeline process. The result is highly efficient ML models created by non-experts with minimal coding.

The application of Auto ML tools in different industries like hospitalities (Baharun, et al., 2022, p. 17), manufacturing (Xiao, et al., 2024, p. 9), medicine (el Ariss, et al., 2024, p. 141), ecology (Prasad, et al., 2021, p. 1), agriculture (Malounas, et al., 2024, p. 2), education (Sulova, 2024), etc.

has been explored by various researchers. In some cases, researchers compare the results of Auto ML against conventional ML depending on the accuracy of the models. Some of them have realized the model trained with Auto ML is more accurate than if it has trained with conventional ML (Prasad, et al., 2021, p. 12). Comparison between trained models with different Auto ML tools has been also provided (Xiao, et al., 2024, p. 11), (Opara, E., Wimmer, H. and Rebman, 2022, p. 1).

1. AUTOMATED MACHINE LEARNING

The definition of AutoML is not unambiguously defined in terms of the degree of automation. Some authors believe that the full automation of the process is necessary to define the concept of AutoML (Xanthopoulos, et al., 2020). Others, in addition to full automation, define the concept of "man in the loop" (Wang, et al., 2021, p. 3), related to partial automation and human participation in the process. In this regard, software vendors provide varying degrees of automation through the Auto ML tools offered. Fully automated processes are characterized by a lack of need for data science experts, and this leads to the democratization of Auto ML (Xanthopoulos, et al., 2020), because such models are focused on business power users. An essential feature of fully automated ML applications, however, is that the result is individual solutions that are less transparent. In most cases, the user cannot understand the output visualizations, which leads to the need for Auto ML to indicate which models have been viewed and why. Another feature is the preset ML algorithms, which are a finite number for solving a certain problem, for example, classification. Provided that the user wishes to use a different algorithm, this is usually not possible.

Results of a study on the performance of Auto ML tools, compared to the use of conventional ML have been published by (Xiao, et al., 2024, p. 11). The study assesses the toxicity of nanomaterials in the production of nanoproducts. In training with conventional tools, algorithms RF (Random Forest), SVM (Support vector machine), GBT (Gradient boosted trees) are used. Compared to them are AutoML tools are Vertex AI, Azure, Dataiku. The models are compared by evaluation metrics Accuracy, F1 score, Precision, Recall. In terms of Accuracy, Auto ML models give an accuracy of 0.97 (Vertex AI), versus 0.95 with conventional ML (GBT). In terms of, Precision has the predominance of conventional ML: 0.93 (Vertex AI) vs 0.95 (RF).

There are other factors favouring the choice of AutoML tools for creating models. These may include automating the selection of variables involved in the model, which helps speed up training times and improve model accuracy. (Solorio-Fernández, S., Carrasco-Ochoa, J. and Martínez-Trinidad, J., 2022) function engineering, consisting of creating, selecting, and refining variables to improve model performance; setting of hyperparameters (Bartz, et al., 2023). There has been a significant improvement in efficiency in building predictive models using AutoML tools (De Bie, et al., 2022). These functionalities protect the ML process from human error when performing steps, such as skipping balancing the dataset, or adjusting hyperparameters, which would reduce the quality of the model. What has been said so far confirms the expediency of using AutoML tools in the process of building models with machine learning.

2. EVALUATED AUTO ML TOOLS

Our research aims to compare and evaluate the capabilities of leading AutoML tools to solve classification problems. The selection is based on the latest research and classification of tools in the Gartner's Magic Quadrant for Data Science and Machine Learning Platforms for 2024 (Jaffri, et al., 2024). This study includes leading tools such as Azure Automated Machine Learning, Amazon Sage Maker Auto Pilot, H2O AutoML, H2O Flow and Altair AI Studio Auto Model. They

are classified as leaders in the magic quadrant, except for H2O, who is classified as a visionary. Each of these tools includes tools to solve classification problems with ML algorithms.

Amazon Sage Maker Auto Pilot recognizes the type of problem, processes the data, and creates the full set of different complete ML pipelines that are optimized to suggest the list of the customer's potential models. By exposing not only the final models but the way they are trained, meaning the pipelines, it allows users to customize the generated training pipeline, thus catering the need of users with different levels of expertise (Das, et al., 2020). Amazon SageMaker Autopilot independently infers the right kind of forecasting for a specific dataset, including binary classification, multi-class classification, and regression. After that, SageMaker Autopilot exhaustively tests various high-performing algorithms like gradient boosting decision trees, feed forward neural networks, and logistic regression. It provides an explainability report that facilitates an understanding and comprehensive explanation of the models created. (Lenkala, et al., 2023)

Azure AutomatedML is part of Azure Machine Learning workspace and is integrated and depends on Azure infrastructure. It is a cloud-based solution that automates data clean, data label, feature engineering, model training, model evaluation and deploying (Quaranta, et al., 2025, p. 11). It includes features for preprocessing and feature engineering, which automated the transformation of raw data into a machine learning-ready format. by changing the missing values by substituting them with the feature's mean, ensuring that no data points were discarded due to incomplete information (el Ariss, et al., 2024). Azure AutoML proceeds to the model selection and training phase, specifically within the context of a classification task. Utilizing its extensive repository of algorithms AutoML selects a diverse array of classification models that range from traditional methods, such as logistic regression and decision trees, to more complex ones like gradient boosting, SVM and LightGBM. It then applies these models within a robust validation framework, employing a 5-fold cross-validation technique to rigorously evaluate model performance. (el Ariss, et al., 2024). Azure Auto ML supports fully automated ML pipeline with no coding, that is very easy to use for nonexperts. The algorithms and parameters of the created models can be exported so this ensures the transparency of the models. The software supports excellent model explanation and that is a big benefit for the customers.

Altair Ai Studio Auto Model addresses enterprise-grade AI by strengthening the convergence of AI, Internet of Things (IoT), and High Performance Computing (HPC) through integration with other Altair products (Jaffri, et al., 2024). According to Gartner research (Jaffri, et al., 2024), Altair plans to increase data democratization by investing in providing a conversational interface to create workflows using its proprietary analytics translation language.

H2O AutoML is an open-source scalable machine learning platform, that uses large datasets. It uses the grid search method for hyperparameter tuning also. (Yang, et al., 2022, p. 6). In addition, it offers a simpler and more efficient analysis of model interpretability. The H2O AutoML function automates the pipeline of ML models for a given dataset, including data pre-processing, feature engineering, hyperparameter optimization, performance evaluation and interpretability. The H2O AutoML offers a variety of ML models that include supervised learning models, unsupervised learning models and deep learning models. Compared to other AutoML models, H2O AutoML is easier to install and shows high prediction performance (Luo, et al., 2023). It Offers API, incl. Python API, R, Java, Scala, REST and others.

H2O Flow is a Web UI for H2O.ai on top of REST API. It allows users don't need coding, but steps must be configured by users. Users set training parameters like balance classes, max run

time_sec, number of models, algorithms – not fully automated. It supports different ML algorithms and stacking and voting ensembles.

3. EXPERIMENT RESULTS

To evaluate the performance of selected automated ML tools we chose Telco dataset provided by IBM (IBM, 2024). This dataset can be used to train binary classification machine learning models to predict customer churn. To explore and compare auto ML tools, we provided the dataset unaltered with no preprocessing, feature engineering, nor data normalization. The only condition related to the training process was a set time limit for training of maximum 3 hours. The purpose of this experiment was to evaluate the performance of auto ML tools from the point of view of power business users who usually don't have a previous knowledge about data preprocessing, preparation, feature engineering or other techniques and approaches from machine learning pipeline. Trained models are ranked according to Area Under Curve (AUC), which gives a comprehensive view on model's predictive power. The best performing model from every auto ML tool is then selected and scored against one and the same test dataset to ensure comparable results.

AutoML tools train multiple models using different machine learning algorithms. During the training different strategies for hyperparameter optimization as used. Trained models are evaluated and ranked according to the chosen measures which in our practical experiment is set to AUC. To control the training process and prevent the extensive usage of computational resources we set training limits to maximum number of models 20 or time limits of 3 hours. As a result, Azure AutomatedML produced 59 models, Azure SageMaker – 20 models set as MaxCandidate parameter, H2O AutoML API and Flow – 20 + 2 ensembles and Altair AI AutoML – 9. Used algorithms include decision tree bases homogenic ensembles like Random Forest, Gradient Boosting Machines, XGBoost, Distributed Random Forest, neural networks like Deep Learning feed forward network, others like Logistic Regression, Generalized Linear Models, Naïve Bayes and others. Auto ML tools also trained heterogenic ensembles like Stacking and Voting ensembles. The only tool that doesn't support ensembles is Altair AI Studio Auto Model.

The ranking of the trained models is done automatically by the respective auto ML tool according to the chosen metrics. The result is a model leaderboard with description of the model – used algorithm, data normalization approach or hyperparameters. Amazon SageMaker was the only auto ML tool with lack of transparency. From the produced model leaderboard it's not possible to determine the used algorithm, nor other important features of the models.

The model leaderboard with top 5 of each auto ML models ranked according to the AUC is shown in figure 1. Evident from the figure, AUC varies from 0.7053 (Altair) to 0.8560 (Amazon SageMaker). The best performing models are ensemble models built with Amazon SageMaker with highest AUC of 0.8550-0.8560.

The second cohort of models ranked according to the AUC is formed by stacked ensemble models trained with H2O AutoML with Python API with AUC of 0.8517 and 0.8505 respectively. Model trained with algorithms Gradient Boosting Machines and General Linear Models in the same auto ML tool have very close to these values of AUC of 0.8486 and 0.8484.

Ensembles trained in Azure AutomatedML are ranked in the middle with Voting ensemble's AUC of 0.8480 and Stacked ensemble's AUC of 0.8477. H2O Auto ML models trained with H2O Web Flow interface have similar performance with AUC of 0.8478 and 0.8475. The worst performing models from all evaluated auto ML were those trained with Altair AI Studio Auto Model. Their AUC varies from 0.7053 to 0.8458.

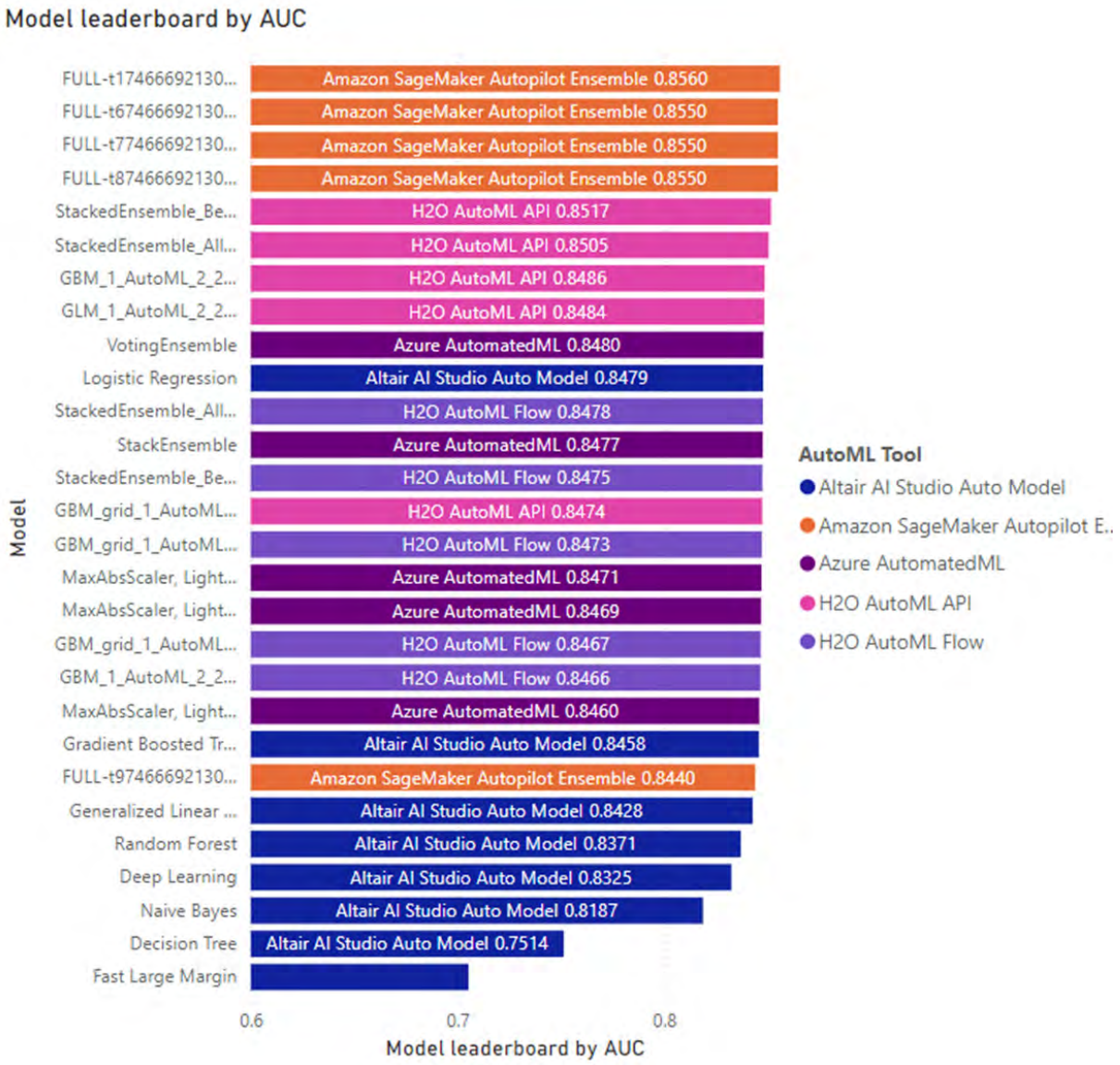

Figure 1. Model leaderboard of top 5 models of each auto ML tools (developed by authors)

On the second stage of evaluation comparison, we selected the best models from each auto ML tool and applied them on one and the same test dataset. The test dataset wasn't used during the training process to ensure that all results are comparable. The performance evaluation included analysis on the following metrics – specificity, sensitivity, F1 score and weighted accuracy. Results from the models' scoring are included in table 1. The best model by each metric is highlighted. All models included in this table are trained with heterogenic ensemble algorithms like Stacked or Voting Ensemble. The only single classifier is the best from Altair AI Studio models, trained with Logistic Regression. At the time of our research, Auto Model in Altair Studio didn't support training ensemble models.

As shown in table 1 there is no clear winner in the evaluation comparison between models. The model with the highest specificity, i.e. the ability to correctly negative class (non-churners), is the Azure AutomatedML ensemble with specificity of 0.925. At the same time the sensitivity of this model is the lowest with value of 0.513. This is due to the fact that we used the default training settings and didn't specify a strategy for class balance. With strongly imbalanced datasets like the one used in this research it's expected that trained models are more fitted to the majority class. Similarly, the model trained in Altair AI Studio has significant difference between it's specificity and sensitivity in favor of the first.

The model with the best ability to correctly classify the positive class (churners) is H2O AutoML model trained with Python API. Its sensitivity is 0.715. Because of the class balance applied during the training process, this model shows a low difference between its specificity and sensitivity. A class balance strategy has been automatically applied also by Amazon SageMaker Autopilot Ensemble model and the performance on both classes is quite similar to the results of the H2O AutoML model.

Table 1

Performance of the best models from each Auto ML tool

AutoML Tool	Specificity	Sensitivity	F1 score	Weighted Accuracy
Amazon SageMaker Autopilot Ensemble	0.883	0.692	0.696	0.829
H2O AutoML API Stacked Ensemble	0.801	0.715	0.633	0.785
Azure AutomatedML Voting Ensemble	0.925	0.513	0.601	0.799
Altair AI Studio Auto Model LogReg	0.871	0.589	0.605	0.794

Source: Own elaboration

In order to determine the best model from those presented in table 1, we suggest using complex metrics assessing the predictive power towards both classes. Such metrics are F1 score and weighted accuracy. The metrics F1 score is calculated as a harmonic means of precision and recall, thus taking into account both the accuracy of positive predictions and the sensitivity of the model. The weighted accuracy metric uses weights to accommodate the class imbalance in the training dataset. The model with the highest F1 score and weighted accuracy is Amazon SageMaker Autopilot ensemble with F1 score of 0.696 and weighted accuracy of 0.829. Due to its comprehensive predictive power, we can determine this model as the best one from those evaluated in the research.

CONCLUSION

Auto ML tools have significantly advanced through the years, improving the efficiency of predictive modelling by allowing the creation of high quality models with minimal user interaction. These tools are highly efficient in terms of time since they facilitate the development and deployment of models which a big plus for organizations with limited resources. In addition, the no-code or low-code approach that is applied to many AutoML platforms makes machine learning easily accessible for those who couldn't have a deep knowledge and experience of the process.

One of the most important findings of this study is the fact that ensemble methods are more effective than single classifiers. In majority of the cases, ensemble techniques give better accuracy, sensitivity and specificity when compared to single classifiers. Among all the evaluated tools, Amazon SageMaker Autopilot appeared to be slightly better, thanks to ensemble techniques that provided better performance. Another two auto ML tools, H2O Auto ML and Azure Automated ML, also performed well with slightly lower performance metrics to Amazon SageMaker Autopilot. Of all the evaluated tools, Altair AI Studio was placed last, implying that its model performance and features could still be optimized. This is also the tool that doesn't implement ensembles with its Auto Model feature which can explain the big difference with other models.

Future research will be oriented to extend the evaluation of auto ML tools with the addition of other tools and platforms like Google Cloud AutoML, DataRobot, Auto-sklearn, Auto-Weks, TPOT and others. The scope of this future analysis will be broadened to include a comprehensive analysis of their explainable features which enhance the model transparency. Such functionality is essential to applications that require decision explanations and accountability, especially in regulated industries.

The performance of AutoML tools may vary depending on the type of predictive task, for example classification, regression, computer vision and text analytics. Majority of auto ML tools provide no-code or low-code functionality to train predictive models solving these tasks and future research will be conducted to evaluate and compare the performance of auto ML tools regarding solving different predictive tasks.

REFERENCES

1. BAHARUN, N. et al. (2022) Auto Modelling for Machine Learning: A Comparison Implementation between RapidMiner and Python. International Journal of Emerging Technology and Advanced Engineering. 12(5). pp. 15-27.
2. BARTZ, E. et al. (2023) Hyperparameter Tuning for Machine and Deep Learning with R: A Practical Guide. 1st ed. s.l.:Springer.
3. CONSUEGRA-AYALA, J. et al. (2022) Intelligent ensembling of auto-ML system outputs for solving classification problems. Information Sciences. Volume 609. pp. 766-780.
4. DAS, P. et al. (2020) Amazon SageMaker Autopilot: a white box AutoML solution at scale. s.l.. s.n.. pp. 1-7.
5. DE BIE, T. et al. (2022) Automating data science. Communications of the ACM. 65(3).
6. EL ARISS, A. et al. (2024) Development and validation of a machine learning framework for improved resource allocation in the emergency department. American Journal of Emergency Medicine. Volume 84. pp. 141-148.
7. IBM (2024) Telco customer churn. [Online] Available at: https://www.ibm.com/docs/en/cognos-analytics/12.0.0?topic=samples-telco-customer-churn [Accessed 15 11 2024].
8. JAFFRI, A. et al. (2024) Gartner. [Online] Available at: https://www.gartner.com/en/documents/5509595 [Accessed 2 10 2024].
9. LENKALA, S. et al. (2023) Comparison of Automated Machine Learning (AutoML) Tools for. Computers. 12(10). p. 197.
10. LUO, J. et al. (2023) Prediction of biological nutrients removal in full-scale wastewater treatment plants using H2O automated machine learning and back propagation artificial neural network model: Optimization and comparison. Bioresource Technology. 390(129842).
11. MALOUNAS, I. et al. (2024) Early detection of broccoli drought acclimation/stress in agricultural environments utilizing proximal hyperspectral imaging and AutoML. Smart Agricultural Technology. Volume 8.
12. OPARA, E., WIMMER, H. AND REBMAN, C. (2022) Auto-ML Cyber Security Data Analysis Using Google. Azure and IBM Cloud Platforms. Prague, Institute of Electrical and Electronics Engineers Inc.
13. PRASAD, D. et al. (2021) Automating water quality analysis using ML and auto ML techniques. Environmental Research. Volume 202. pp. 1-14.
14. QUARANTA, L. et al. (2025). A multivocal literature review on the benefits and limitations of industry-leading AutoML tools. Information and Software Technology. 178(107608).
15. SOLORIO-FERNÁNDEZ, S., CARRASCO-OCHOA, J. AND MARTÍNEZ-TRINIDAD, J. (2022) A survey on feature selection methods for mixed data. Artificial Intelligence Review. Volume 55. p. 2821–2846.
16. SULOVA, S.(2024) Application of Natural Language Processing Technologies to Improve Accessibility of E-Learning Resources. Varna. Scopus. pp. 180-184.

17. WANG, D. et al. (2021) AutoDS: Towards Human-Centered Automation of Data Science. s.l., s.n.. pp. 1-12.
18. XANTHOPOULOS, I. et al. (2020) Putting_the_Human_Back_in_the_AutoML_Loop. Copenhagen, s.n.
19. XIAO, X. et al. (2024) Automated machine learning in nanotoxicity assessment: A comparative study of predictive model performance. Computational and Structural Biotechnology Journal. Volume 25. pp. 9-19.
20. YANG, H., et al. (2022) Prediction of Wave Energy Flux in the Bohai Sea through Automated Machine Learning. Journal of Marine Science and Engineering. 10(8).