# OVERVIEW ON AUTOML TECHNOLOGIES

## 1. xTech Presentation

xTech is a Bip Center of Excellence born in 2013 and specialized in two technological fields, Data & Decision Sciences and Digital Platforms. Since Its birth xTech has seen a constant and significant growth, and today it consists of more than 170 professionals with over 100 clients. Within Data & Decision Sciences area, xTech has created The largest professional Data Scientist community in Italy, 70 in total.

## 2. Introduction

Data Science is an interdisciplinary subject that embraces maths, computer science and domain knowledge to extract maximum value from data and generate insights that can drive decision-making processes. Data Science framework is represented by "cross-industry standard process for data mining" model, known as CRISP-DM.

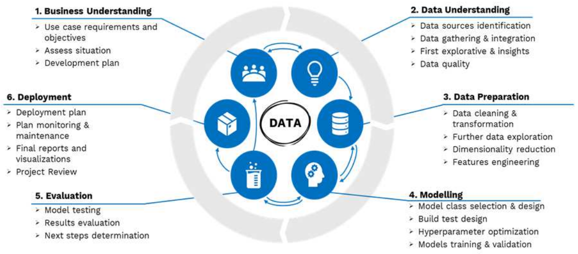

1. Business Understanding
Use case requirements and objectives
Assess situation
Development plan
2. Data Understanding
Data sources identification
Data gathering & integration
First explorative & insights
Data quality
3. Data Preparation
Data cleaning & transformation
Further data exploration
Dimensionality reduction
Features engineering
4. Modelling
Model class selection & design
Build test design
Hyperparameter optimization
Models training & validation
5. Evaluation
Model testing
Results evaluation
Next steps determination
6. Deployment
Deployment plan
Plan monitoring & maintenance
Final reports and visualizations
Project Review
DATA

*Figure 1 – CRISP-DM Model*

CRISP-DM represents Data Science framework as an iterative process, that starts from the understanding of the business context and goals and ends with the deployment of the technical solution. In the modeling stage, Data Science highly leverages on Machine Learning, a discipline that study and develops mathematical-statistical models that learn patterns from data and use them to make predictions.

Automated Machine Learning ("AutoML") is the process of automatically building and tuning a complex Machine Learning model.

Automated Machine Learning has been developed as an answer both for democratizing advance techniques by enabling business analysts to act as data scientists and for automatizing repetitive processes for those professionals already expert. Our study starts from this thesis and investigates if current AutoML technologies can actually replace Data Scientists in any activities reported in CRISP-DM.

## 3. Automated Machine Learning and Data Scientists

Figure 2 represents CRISP-DM activities and AutoML impacts on them, by distinguishing between Expert Data Scientist and Citizen Data Scientist. Expert Data Scientist is a data scientist with a solid background in Data Science & Artificial Intelligence, while a citizen Data Scientist is a business analyst whose role involves data-driven tasks.

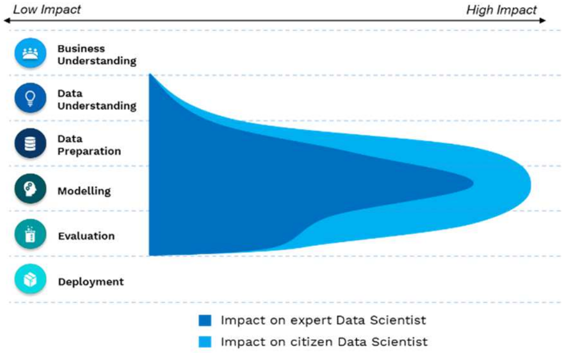


*Figure 2 – Automated Machine Learning impacts*

AutoML impact increases accordingly to the degree of automatization of each activity: understanding stages highly need human comprehension and problem solving skills, while model class selection and evaluation are more algorithmic and thus easier to automate. At the same time AutoML technologies influences more citizen than expert DS, due to the lack of specific competencies of the formers.

## 4. Market Overview

AutoML is still a niche topic but nowadays an increasing number of companies is adopting these technologies.
From a supply perspective, we conducted a broad market scouting to identify main vendors and products of AutoML, leveraging on
Gartner and Forrester reports (end of 2018), and on our expertise in the industry.

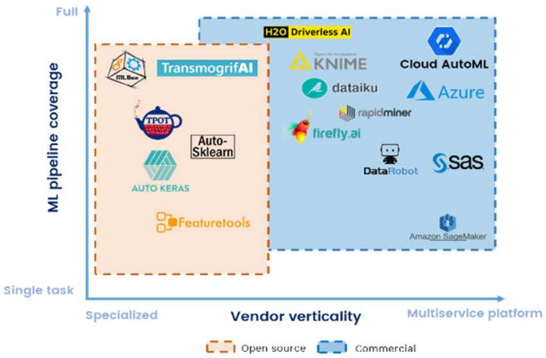


*Figure 3 - Vendors and Products*

As Figure 2 reports, we tried to classify products according to two dimension: Vendor verticality and Machine Learning (ML) pipeline coverage. Vendor verticality refers to provider offering breadth, while ML pipeline coverage indicates the product capability of automatizing different activities involved in the creation of a Machine Learning model, from data cleaning to model validation. Big vendors tend to propose a multi-purpose offering, while open source technologies are strongly vertical on ML and AutoML solutions.

## 5. Products Testing

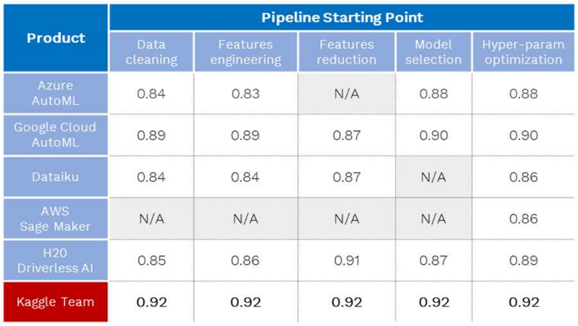
We executed a testing phase where we tried different products on different datasets. We chose the following datasets:

- Multinational Bank: Kaggle dataset of binary classification (Santander Competition)
- Energy Player: Bip dataset of time-series regression referred to an energy commodity
- Fishing Market: Bip dataset of time-series regression referred to European countries fishing

AutoML products generally automatize the following activities: data cleaning, features engineering, features selection, model selection and hyper-parameter optimization.

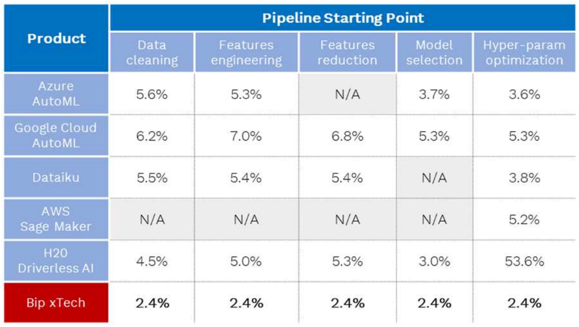
In order to detect from which stage they could over-perform Data Scientists, we adopted an iterative approach, switching progressively the pipeline starting point to the subsequent activity. Consistently, the input dataset is the one prepared up to the preceding stage by reference team. Of course, if a tool did not support the automation of an activity, it wasn't tested on this. Finally, performances were evaluated with respect to results achieved by best-performer Data Scientists teams.

The following tables show tests results.



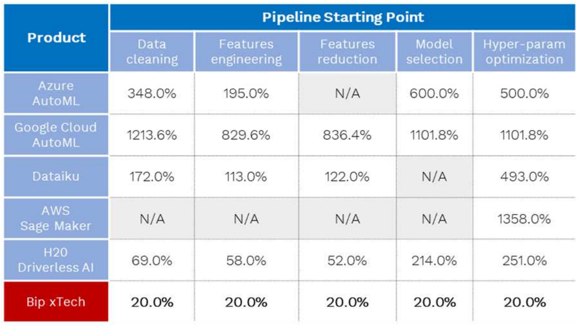
| Product | Pipeline Starting Point | | | | |
|---|---|---|---|---|---|
| | Data cleaning | Features engineering | Features reduction | Model selection | Hyper-param optimization |
| Azure AutoML | 0.84 | 0.83 | N/A | 0.88 | 0.88 |
| Google Cloud AutoML | 0.89 | 0.89 | 0.87 | 0.90 | 0.90 |
| Dataiku | 0.84 | 0.84 | 0.87 | N/A | 0.86 |
| AWS Sage Maker | N/A | N/A | N/A | N/A | 0.86 |
| H20 Driverless AI | 0.85 | 0.86 | 0.91 | 0.87 | 0.89 |
| Kaggle Team | 0.92 | 0.92 | 0.92 | 0.92 | 0.92 |

*Figure 4 - Multinational Bank results*



| Product | Pipeline Starting Point | | | | |
|---|---|---|---|---|---|
| | Data cleaning | Features engineering | Features reduction | Model selection | Hyper-param optimization |
| Azure AutoML | 5.6% | 5.3% | N/A | 3.7% | 3.6% |
| Google Cloud AutoML | 6.2% | 7.0% | 6.8% | 5.3% | 5.3% |
| Dataiku | 5.5% | 5.4% | 5.4% | N/A | 3.8% |
| AWS Sage Maker | N/A | N/A | N/A | N/A | 5.2% |
| H20 Driverless AI | 4.5% | 5.0% | 5.3% | 3.0% | 53.6% |
| Bip xTech | 2.4% | 2.4% | 2.4% | 2.4% | 2.4% |

*Figure 5 - Energy Player results*



| Product | Pipeline Starting Point | | | | |
|---|---|---|---|---|---|
| | Data cleaning | Features engineering | Features reduction | Model selection | Hyper-param optimization |
| Azure AutoML | 348.0% | 195.0% | N/A | 600.0% | 500.0% |
| Google Cloud AutoML | 1213.6% | 829.6% | 836.4% | 1101.8% | 1101.8% |
| Dataiku | 172.0% | 113.0% | 122.0% | N/A | 493.0% |
| AWS Sage Maker | N/A | N/A | N/A | N/A | 1358.0% |
| H20 Driverless AI | 69.0% | 58.0% | 52.0% | 214.0% | 251.0% |
| Bip xTech | 20.0% | 20.0% | 20.0% | 20.0% | 20.0% |

*Figure 6 - Fishing Market results*

## 6. Conclusions and next steps

Tests results highlights benefits and drawbacks of Automated Machine Learning. AutoML tools surely provide a good baseline performance with less effort than a traditional Machine Learning approach. Going into detail of numbers, these products show to be mostly effective when implemented in the final stages of the Machine Learning pipeline, right after features engineering. At the same time, we should report that AutoML performs better when the problem shows a simple data structure, while it suffers when high domain knowledge and diverse data are involved. This is clearly demonstrated by Fishing Market results, where the tools are not able to understand the hierarchical structure of the dataset. Moreover, AutoML can intervene only when a first version of the plain dataset is ready. Finally, results clearly show that custom Machine Learning still remains the approach that leads to the best achievable performances even on equal footing.

Next steps and developments of this study are the continuation of products testing, with a particular focus on open source solutions, and the analysis of the "performance-time efficiency" trade-off between traditional Machine Learning and Automated Machine Learning.