3rd Conference on Production Systems and Logistics

# The Potential Of AutoML For Demand Forecasting

Kathrin Julia Kramer[1], Niclas Behn[2], Matthias Schmidt[1]

[1]*Institute of Product and Process Innovation (PPI)/ Leuphana University, Lueneburg, Germany*
[2]*Schaeffler Technologies AG & Co. KG, Schweinfurt, Germany*

## Abstract

In demand forecasting, which can depend on various internal and external factors, machine learning (ML) methods can capture complex patterns and enable precise forecasts. Accurate forecasts facilitate targeted, demand-oriented planning and control of production and underline the importance of this task. The implementation of ML-algorithms requires knowledge of the specific domain as well as knowledge of data science and involves an elaborate set up process. This often makes the application of ML to potential industrial problems economically unattractive. The major skills shortage in the field of data science further exacerbates this. Automation and better accessibility of ML methods is therefore a key prerequisite for widespread use. This is where the principle of automated ML (AutoML) comes in, automating large parts of a ML pipeline and thus leading to a reduction in human labour input. Therefore, the aim of the publication is to investigate the extent to which AutoML solutions can generate added value for demand planning in the context of production planning and control. For this purpose, publicly available datasets deriving from Walmart as well as an anonymised manufacturing company are used for short-term and long-term forecasting. The AutoML tools from Microsoft, Dataiku and Google conduct these forecasts. Statistical models serve as benchmarks. The results show that the forecasting quality varies depending on the software, the input data and their demand patterns. Overall, the prepared models from Microsoft show the most accurate results in average and the potential of AutoML becomes particularly clear in the short-term forecast. This paper enriches the research field through its broad application, giving valuable insights into the use of AutoML tools for demand planning. The resulting understanding of limitations and benefits of AutoML tools for the case studies presented fosters their suitable application in practice.

## Keywords

AutoML; Demand Forecasting; Sales Forecast; Machine Learning; Manufacturing; Production Planning

## 1. Introduction

ML-based demand forecasting offers the possibility to reflect various influencing parameters (e.g. currency exchange rates, sales region) and use those to identify complex non-linear patterns for forecasting future demand [1]. Accurate predictions enable an adequate planning of the production and procurement processes leading to less waste of resources (material, labour, capital) [2]. To foster the widespread use of machine learning (ML) in an industrial context, automated machine learning (AutoML) gains in importance as it aims to reduce the required knowledge in data science and time spend to set up a ML model [3–5]. Thus, it presents a possible solution for overcoming the skill shortage in the field of data science, which is currently one of the major barriers for the application of ML [2,6,7]. In addition, the shorter development time achieved by automating parts of a ML pipeline makes ML solutions more economically attractive [4,8]. Thus, researchers

DOI: https://doi.org/10.15488/12162

investigated the benefit of AutoML solutions in different case studies across various use cases (e.g. other production planning and control (PPC) tasks [9] or healthcare [10]). However, studies in regards to demand planning [11–13] in the context of PPC [14] remains a research gap. This paper aims at closing this gap by answering the research question: "*Can AutoML solutions support demand forecasting?*". The research question is embedded in the current state of research by highlighting the results of topic-related case studies. The third section presents the research methodology for achieving the objective of this paper. A generally valid answer to the research question is not the aim of this work, although it should be possible to make an assessment as broad as possible. To facilitate a comprehensive assessment, three AutoML tools in two production environments for forecasting demand of different product groups as well as for a short-term and long-term horizon are tested. Broad application enables a comprehensive evaluation of the prediction accuracy of the chosen AutoML tools and allows first conclusions for the use of AutoML solutions in demand forecasting. The fourth section presents the results of the research methodology used. The final section of the paper draws a conclusion and presents a future research agenda.

In summary, this paper enhances the current research through its broad application and comprehensive evaluation and thus allows conclusions on the potential use of AutoML tools for demand planning. In particular, an automated approach is required for scaling ML [8], i.e. ML-based prediction across different levels of observation (e.g. total, product group, product) and on a horizontal axis (e.g. each product).

## 2. Current state of research

This section explains the concept of AutoML and presents existing case studies that deal with AutoML in demand planning.

The herein used definition of AutoML was first introduced in 2014 and foresees an automation across the ML-pipeline [15]. The concept of AutoML has the objective to reduce the outlay of data scientists for ML projects and should instead enable domain experts to use ML methods without high level of statistical and ML knowledge [16]. More precisely, Yao et al. define AutoML as a combination of automation and ML and understand AutoML as an automated setup of a ML-pipeline with limited computing power and limited (or no) human support [17]. A common ML-pipeline consists of business and data understanding, data preparation, modelling, evaluation and deployment tasks, e.g. used in the industry-independent Cross Industry Standard Process for Data Mining (CRISP-DM) [18]. This framework is already used for demand forecasting [19–21]. The nature of the process involves feedback loops and continuous readjustments of assumptions, forming a life cycle process [18]. In particular, the processes of data understanding, data preparation, modelling and evaluation require the expertise of data scientists and could therefore be automated [5]: data preparation foresees to select features, clean and transform those as well as generate new features. The modelling process contains the choice of an algorithm and the optimization of its hyper-parameters and for artificial neural networks (ANN) also the definition of the net architecture [18,5]. Various facets can be investigated during models' evaluation, whereby the prediction accuracy is usually the focus [22]. The evaluation can take place when the algorithm converges during training or in order to save time as well as computing power, for instance a predefined budget of computing resources can be used as stop criteria [5].

For demand planning, AutoML has already been used in several studies. The focus of these studies is primarily in the sales and marketing environment [11–13]. The study by Gonçalves et al. is the only contribution that investigates AutoML for demand planning in the context of PPC [14]. Ford et al. apply a self-developed AutoML solution for the products of an alcoholic beverage distributor. For this purpose, they generate forecasts for three horizons, 1 month, 3 months and 12 months, and include autoregressive methods and the average for a comparison. They use two univariate datasets and mean squared error as a quality criterion. The AutoML models achieve worse results, which are below the forecasting quality of the autoregressive methods. For one of the datasets, which is characterised by a high proportion of noise, the average method achieves the best results [12]. Henzel and Sikora use AutoML to create ANN and compare them with manually created XGBoost and ANN models. They use a dataset of everyday consumer goods and focus on the impact of promotions on sales. The models created with AutoML are superior to both manually created models in 2 of 12 cases (according to root mean squared error (RMSE) and mean absolute error (MAE)), and better than the manually created ANN in all cases [13]. Dai and Huang create a Long Short Term Memory (LSTM) model with a special loss function, which they optimise with hyper-parameter search. For comparison, they create six ML models with AutoML, which they use as benchmarks. They use a sparse consumer goods dataset that contains causal information of sales (e.g. holidays, promotions). The authors use weekly and monthly data to forecast cumulative sales at the store level. The LSTM model achieves better results (according to mean absolute percentage error (MAPE) and root mean squared percentage error) compared to the AutoML models, whereby the performance of the LSTM model decreases with increasing forecast horizon [11]. Gonçalves et al. compare statistical (Naïve, exponential smoothing, ARIMA, ARIMAX) and ML models (feed-forward ANN, random forest, support vector regression, recurrent ANN), including a model with AutoML. They forecast the demand for electronic components of a manufacturer from the automotive supply sector and follow a multivariate approach with various leading indicators. In particular, they investigate whether and to what extent a multivariate approach is superior to a univariate approach in different phases of the product life cycle. In the quality criterion normalized MAE (nMAE) used, AutoML achieves the fourth best performance (out of 9) on average across all phases of the product life cycle, with a considerable gap between it and the following statistical methods: Naïve, exponential smoothing and ARIMA [14].

The studies presented show that AutoML models tend to perform worse than manually created ML models and better than statistical models. This is particularly the case with multivariate forecasts. For univariate forecasts and one-step forecasts, statistical methods tend to perform better. With the few identified studies, the described area of research is still underrepresented. Thus, further investigations in the field of demand forecasting, and especially in the context of PPC are necessary. To add to this note, continuous progress in the field of AutoML makes results of past studies hard to interpret for assessing the potential of current AutoML solutions. In addition, existing studies on demand forecasting have not yet compared several AutoML solutions. Thus, this paper contributes to the existing research by conducting a comparison of several AutoML solutions with statistical methods for the area of demand forecasting in the context of PPC.

## 3. Research methodology

The research methodology of this paper is of empirical nature. A transparent setup of different experiments enables an in-depth understanding of the limitations and benefits when using the chosen AutoML tools and facilitates researchers to transfer this methodology to different case studies as well as AutoML tools.

To begin, the authors identified 31 existing AutoML tools. The pool of potential tools reduces to eight tools as only those tools fulfil the following criteria: they offer a test version or academic licence, can handle time series data and offer at least partial automatic data preparation. Of these, three tools are chosen as examples for the investigations in this paper: the AutoML solutions Microsoft Azure Automated ML, Google Cloud AutoML Tables and Dataiku Data Science Studio. To emphasise again, a selection of more than one tool is of importance to understand possible deviations across the tools. An aspect that was so far not analysed. ARIMA and exponential smoothing are taken as benchmarks as these methods are most commonly used in practice and do not require extensive data science knowledge [23,24]. For the investigation of the chosen tools, the following assumptions are considered with the objective to facilitate a comparable set up as well as test the AutoML solutions in different settings. Figure 1 shows these specifications that are structured according to the CRISP-DM phases (grey boxes).

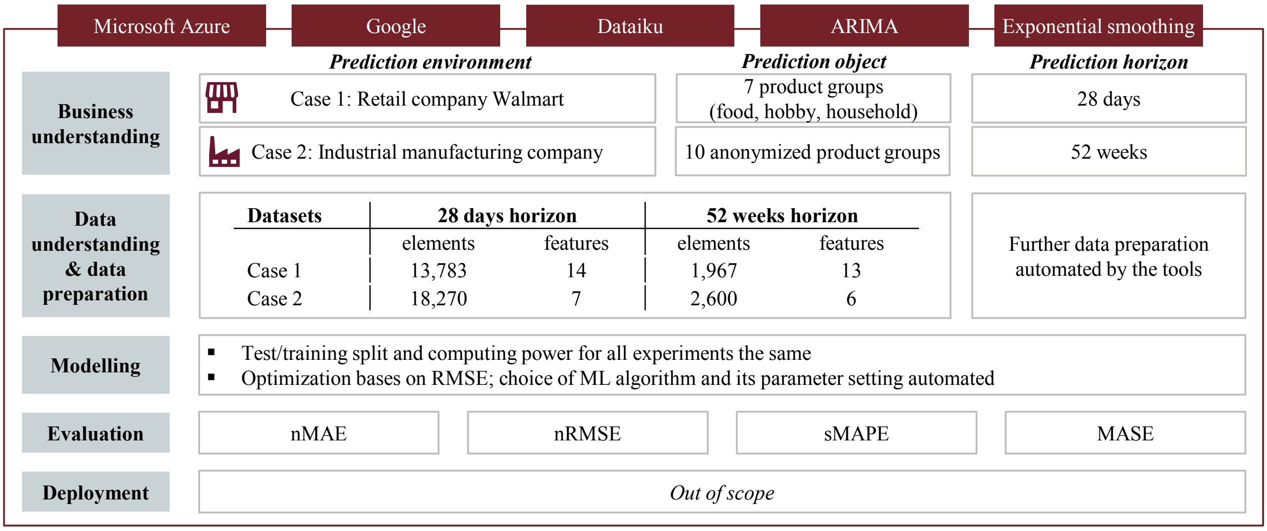
Microsoft Azure
Google
Dataiku
ARIMA
Exponential smoothing
Prediction environment
Prediction object
Prediction horizon
Business understanding
Case 1: Retail company Walmart
7 product groups (food, hobby, household)
28 days
Case 2: Industrial manufacturing company
10 anonymized product groups
52 weeks
Data understanding & data preparation
Datasets
28 days horizon
52 weeks horizon
elements
features
elements
features
Case 1
13,783
14
1,967
13
Case 2
18,270
7
2,600
6
Further data preparation automated by the tools
Modelling
Test/training split and computing power for all experiments the same
Optimization bases on RMSE; choice of ML algorithm and its parameter setting automated
Evaluation
nMAE
nRMSE
sMAPE
MASE
Deployment
Out of scope

Figure 1: Specifications of examinations

The demand forecasts take place for two different companies: the retail goods company Walmart (case 1) and an anonymized company that manufactures industrial goods (case 2). In order to compare the forecast quality at different forecast horizons, a yearly and monthly forecast will be prepared [25]. 28 days and 52 weeks are chosen as prediction horizons to analyse whether there are performance differences between a short- and long-term perspective. The aim is to obtain the total demand per product group (label attribute) in order to support decision-making processes, e.g. the decision on long-term supplier contracts and the acquisition of a new production plant from a long-term perspective or the timing of the procurement of bought-in parts and production from a medium-term perspective. To investigate how robust the AutoML solutions predict, it is of importance to test the tools on different demand patterns. This is the case for the chosen product groups, seven product groups for case 1 from the food, hobby and household sectors and ten anonymized product groups for case 2. The datasets are publicly available [26,27]. As the PPC is the focus of this work, the public datasets have been modified to remove the store/ warehouse levels. The datasets consist of 13,783 elements (case 1) and 18,270 elements (case 2) of daily demand per product group for the prediction horizon of 28 days. For the prediction horizon of 52 weeks, case 1 has 1,967 elements and case 2 counts 2,600 elements of weekly demand per product group. Table 1 gives an overview of the different groups of features. Three main differences exist regarding the features. First, the dataset of case 1 counts more features that are descriptive then case 2. It contains additional information on the occurrence of public events and governmental support schemes of three US states. The dataset of case 2 lacks these features and only consists of the label attribute and the date. To test the tools on a multivariate setting, further time information (e.g. weekday, calendar week, year) are added. Secondly, when transforming the datasets from daily to weekly data, some features are excluded (e.g. weekday), transformed (e.g. event features) or added (e.g. calendar week). Thirdly, depending on the AutoML tool, further features are added: the ML tools of Microsoft Azure and Dataiku add information on school and bank holidays. This addition is only made in case 1, as the region is known, whereas it is unknown in case 2. Besides, Microsoft Azure Automated ML is the only tool that generates features about the seasonal and trend component of the demand time series as well as lag features for public events. Thus, this tool generates most features in comparison to the other two tools.

Table 1: Overview of features

| Initial feature group | Case 1 | | | Case 2 | | |
|---|---|---|---|---|---|---|
| Demand per product group (label) | x | | | x | | |
| Time information (e.g. date, week) | x | | | x | | |
| Public events | x | | | | | |
| Governmental support scheme | x | | | | | |
| **Additional AutoML feature group** | **Azure** | **Dataiku** | **Google** | **Azure** | **Dataiku** | **Google** |
| Bank- and school holidays | x | x | | | | |
| Lag features for public events | x | | | | | |
| Seasonal and trend components | x | | | x | | |
| **Number of reflected feature groups** | **7** | **5** | **4** | **3** | **2** | **2** |

x = part of the dataset Azure = AutoML tool by Microsoft Azure Dataiku = AutoML tool by Dataiku Google = AutoML tool by Google

The dataset of case 2 contained missing values that were replaced by zero to enable a proper use of the tools. Besides that, the selected AutoML tools perform the remaining data preparation (e.g. feature selection, feature generation, data normalization). By means of this paper, the objective is to investigate the performance of the automated ML processes. Thus, no further manual preparation takes place. Modelling bases also on the automated decisions of the tools. However, for facilitating comparable results, the test split corresponds to the forecasting horizon, optimization bases on RMSE and on a constant computing power available for training across all tools. For evaluation of the models, first it is outlined whether AutoML tools are able to create predictions that are better than Naïve forecasts and secondly, if those models perform better than the benchmark of statistical methods. The evaluation metrics nMAE, nRMSE, sMAPE, and mean absolute scaled error (MASE) are selected. MAE is to be used because it is an absolute and scale-based measure that has been used before in similar research projects. RMSE is also frequently used [13,24,28]. This paper uses the normalised version of MAE and RMSE (nMAE and nRMSE) so that a comparison over several time series is possible. The mean of the actual values of the forecast horizon is chosen for normalisation. The symmetrical variant of the mean absolute percentage error sMAPE is used as a percentage quality measure. This is more robust than MAPE and allows the evaluation of zero values. Both quality measures are frequently used in the evaluation of time series [24,29]. Hyndman & Koehler argue that the value of sMAPE can be unstable and instead recommend the use of MASE [30]. Thus, the last measure used is the relative quality measure MASE, which has been applied in several studies [23,24,30,31]. The Naïve method functions as a benchmark model for MASE. If the calculated value is below 1 the forecast is better than a Naïve forecast and vice versa [30]. The final phase, the deployment of models, exceeds the scope of this paper.

Overall, the research methodology enables a comprehensive analysis of the potential from AutoML tools with regard to the prediction accuracy for the chosen use cases. By looking at different datasets with various demand patterns and different input features as well as multiple evaluation metrics, this paper contributes to the research field. The results help to understand the limitations and potentials of the observed tools in the investigated environments and lead to hypotheses for future applications.

## 4. Results

This section presents the results of the applied research methodology. The results relate to the best performing and thus chosen model of each AutoML tool as well as ARIMA and exponential smoothing. The training of the respective models took 42 minutes to 100.2 minutes. The first part of the analysis is to check whether the prediction accuracy of the best performing model according to MASE per product group of case 1 and case 2 is above or below 1 for both prediction horizons. In case 1, the predictions of at least one model across all product groups are better than a Naïve forecast. However, in case 2, the provided predictions for product group 001, 007 and 011 for the 52-weeks horizon do not show sufficient results as they are as good as a Naïve forecast. As second part of the analysis, Table 2 summarizes the average results of the best AutoML tool in comparison to the best statistical model. For the following comparison, the product groups 001, 007 and 011 are disregarded as they would otherwise distort the picture.

Table 2: Prediction accuracy of the best AutoML tool in comparison to best statistical model

| Case | Horizon | Best AutoML tool | Best statistical method | nMAE | nRMSE | sMAPE | MASE |
|---|---|---|---|---|---|---|---|
| 1 | 28 days | Azure | Exponential smoothing | +17.87% | +18.96% | +17.62% | +23.88% |
| | 52 weeks | Dataiku | ARIMA | +2.92% | +7.99% | +2.11% | -4.46% |
| 2 | 28 days | Azure | ARIMA | +18.81% | +17.81% | +8.05% | +7.11% |
| | 52 weeks | Dataiku* / Azure** | Exponential smoothing | -10.71% | -10.08% | -6.51% | -10.85% |

* according to nMAE & nRMSE ** according to sMAPE & MASE

It shows that the prediction accuracy of AutoML across the different experiments is in average in particular beneficial for the short-term forecast. The best performing AutoML tool in this case is Microsoft Azure's tool. Table 2 displays, depending on the evaluation metric, an average improvement of 17.62% to 23.88% for the first case study and 7.11% to 18.81% for the second case study in comparison to the best performing statistical model. The long-term forecasting reflects mixed results. For case 1, three of four evaluation metrics reflect an improved prediction of AutoML by 2.11% to 7.99% in comparison to a statistical method. However, the MASE value indicates a negative effect of -4.46%. The results of the best performing AutoML model in case 2 show worse results (-6.51% to -10.85%) in contrast to the best performing statistical model. To get a more detailed picture of the prediction accuracy of each AutoML tool as well as the benchmark of statistical models across the different product groups, the following figure illustrates the distribution through boxplots of each model according to MASE.

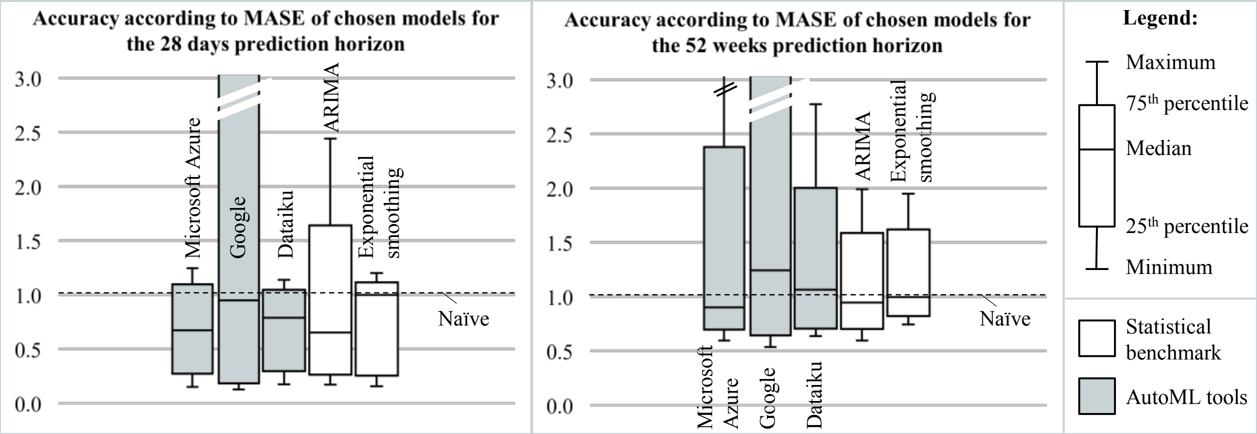

Figure 2: Box plot diagram of prediction accuracies from all product groups per model

It shows that the results of the AutoML tool by Google varies the most. In one case the tool receives the best prediction accuracy (case 1, product group Food 3, prediction horizon of 28 days with MASE of 0.127) and in another case the worst accuracy (case 2, product group Category_001, prediction horizon of 28 days with MASE of 33.885) in relation to all product groups. For the other tools, differences between the two prediction horizons exist. On the short-term horizon, the tool of Dataiku varies the least (MASE between 0.175-1.139), closely followed by exponential smoothing (MASE between 0.156-1.200) and Microsoft Azure's tool (MASE between 0.152-1.246). ARIMA has the lowest median with a MASE value of 0.651, but the second most variation of all values. The tool of Microsoft Azure points out the second lowest median. On the long-term prediction horizon, Figure 2 shows that the prediction accuracy's variation of the statistical models ARIMA and exponential smoothing is lower than the variation from the AutoML models. However, only the median of ARIMA (median MASE of 0.945) and Microsoft Azure (median MASE of 0.901) was lower than one and is thus better than a Naïve forecast. Please refer to Table 3 for an overview of key statistics, also of the other evaluation metrics. When ranking the model according to prediction accuracy, in some cases the ranking differs when looking at different evaluation metrics. For example, the best performing model for the long-term horizon, thus the minimum value of considered metric, is in the case of MAE, nRMSE, sMAPE the ARIMA model and for MASE the tool by Google. However, when looking at the associated mean across all product groups of both case studies, ARIMA is the model in favour.

Table 3: key statistics of investigated models across all product groups of both case studies

| | **Expon. smoothing** | **ARIMA** | **Microsoft Azure** | **Dataiku** | **Google** | **Expon. smoothing** | **ARIMA** | **Microsoft Azure** | **Dataiku** | **Google** | **Expon. smoothing** | **ARIMA** | **Microsoft Azure** | **Dataiku** | **Google** | **Expon. smoothing** | **ARIMA** | **Microsoft Azure** | **Dataiku** | **Google** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **28 days** | MAE | | | | | nRMSE | | | | | sMAPE | | | | | MASE | | | | |
| **Min.** | 0.046 | 0.043 | 0.041 | 0.043 | 0.031 | 0.054 | 0.052 | 0.050 | 0.054 | 0.040 | 0.047 | 0.043 | 0.042 | 0.044 | 0.031 | 0.156 | 0.173 | 0.152 | 0.175 | 0.127 |
| **1st Q.** | 0.058 | 0.059 | 0.065 | 0.063 | 0.071 | 0.081 | 0.081 | 0.079 | 0.085 | 0.085 | 0.056 | 0.057 | 0.069 | 0.062 | 0.075 | 0.352 | 0.355 | 0.390 | 0.418 | 0.240 |
| **Median** | 0.193 | 0.151 | 0.189 | 0.171 | 0.205 | 0.241 | 0.204 | 0.232 | 0.195 | 0.250 | 0.196 | 0.147 | 0.189 | 0.168 | 0.214 | 1.000 | 0.651 | 0.672 | 0.790 | 0.948 |
| **3rd Q.** | 0.311 | 0.324 | 0.294 | 0.295 | 2.969 | 0.402 | 0.408 | 0.344 | 0.373 | 3.646 | 0.324 | 0.353 | 0.315 | 0.313 | 1.874 | 1.029 | 0.837 | 0.952 | 0.952 | 5.463 |
| **Max.** | 1.092 | 0.756 | 0.544 | 0.812 | 22.81 | 1.181 | 0.907 | 0.591 | 0.887 | 29.97 | 0.760 | 0.582 | 0.498 | 0.641 | 2.000 | 1.200 | 2.441 | 1.246 | 1.139 | 33.88 |
| **Mean** | 0.279 | 0.236 | 0.202 | 0.230 | 2.825 | 0.338 | 0.283 | 0.244 | 0.275 | 3.664 | 0.248 | 0.223 | 0.206 | 0.226 | 0.681 | 0.768 | 0.722 | 0.662 | 0.682 | 5.932 |
| **St. Dev.** | 0.294 | 0.218 | 0.147 | 0.204 | 5.983 | 0.325 | 0.249 | 0.171 | 0.228 | 7.868 | 0.219 | 0.184 | 0.149 | 0.186 | 0.845 | 0.391 | 0.560 | 0.394 | 0.338 | 9.933 |

| **52 weeks** | MAE | | | | | nRMSE | | | | | sMAPE | | | | | MASE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Min.** | 0.047 | 0.032 | 0.065 | 0.061 | 0.059 | 0.055 | 0.041 | 0.079 | 0.074 | 0.073 | 0.047 | 0.033 | 0.065 | 0.063 | 0.059 | 0.743 | 0.595 | 0.596 | 0.638 | 0.537 |
| **1st Q.** | 0.099 | 0.094 | 0.093 | 0.074 | 0.107 | 0.123 | 0.123 | 0.112 | 0.095 | 0.131 | 0.101 | 0.096 | 0.093 | 0.075 | 0.110 | 0.897 | 0.810 | 0.796 | 0.773 | 0.750 |
| **Median** | 0.213 | 0.169 | 0.163 | 0.186 | 0.174 | 0.241 | 0.242 | 0.200 | 0.204 | 0.200 | 0.209 | 0.170 | 0.157 | 0.173 | 0.183 | 0.996 | 0.945 | 0.901 | 1.064 | 1.242 |
| **3rd Q.** | 0.259 | 0.321 | 0.227 | 0.295 | 0.858 | 0.309 | 0.364 | 0.266 | 0.398 | 1.187 | 0.243 | 0.290 | 0.225 | 0.287 | 0.730 | 1.290 | 1.185 | 1.102 | 1.230 | 2.427 |
| **Max.** | 0.845 | 0.514 | 1.163 | 0.638 | 4.214 | 0.888 | 0.579 | 1.176 | 0.674 | 5.982 | 0.942 | 0.453 | 0.744 | 0.499 | 1.470 | 1.948 | 1.990 | 3.654 | 2.774 | 17.04 |
| **Mean** | 0.245 | 0.222 | 0.247 | 0.236 | 0.778 | 0.284 | 0.263 | 0.287 | 0.274 | 1.074 | 0.244 | 0.206 | 0.218 | 0.216 | 0.439 | 1.111 | 1.033 | 1.109 | 1.124 | 3.472 |
| **St. Dev.** | 0.214 | 0.153 | 0.275 | 0.177 | 1.300 | 0.234 | 0.171 | 0.283 | 0.191 | 1.838 | 0.226 | 0.131 | 0.187 | 0.144 | 0.493 | 0.345 | 0.367 | 0.699 | 0.527 | 5.300 |

Looking at the short-term prediction in most cases no big differences occur when comparing the average prediction accuracy of AutoML with the statistical benchmark. In 10 out of 17 product groups (58.8%) the average deviation equals to MASE of -0.02-0.18. The remaining seven product groups vary between -0.36 and 11.08. The deviation is even lower when comparing the best performing model. There, only the product group Food 2 and 028 show a deviation of -0.53 and -0.76. Illustrative, the left side of Figure 3 presents the predicted demand of the best performing AutoML model (line-dotted line) and statistical model (circular-dotted line) in relation to the actual demand of product group 028. The deviation of the remaining product groups is only -0.14-+0.14.

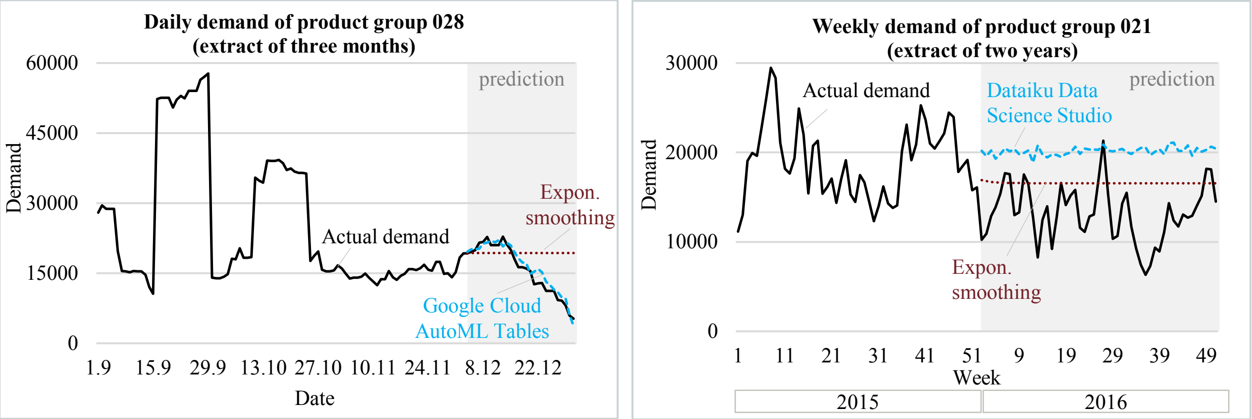

Figure 3: Extract of predicted demand from selected models and actual demand

For the long-term prediction the average deviation between AutoML and the statistical benchmark is in 10 out of 17 product groups between -0.31-+0.25. For the remaining seven product groups, five groups in favour of statistical models (in average 0.50-5.36 more precise than AutoML) and two groups in favour of AutoML (in average 0.68 and 0.85 more precise than the statistical benchmark). When looking at most precise model and not the average, the best performing model of six product groups deviates more than 0.5 between statistical and AutoML models: four in favour for AutoML (0.52-0.60 more precise than the best performing statistical model) and two in favour of the best performing statistical model (0.56 and 0.84 more precise than the best AutoML model). Product group 021, shown in Figure 3, corresponds to the biggest deviation in favour of the statistical benchmark, when looking at the MASE values. Thus, Figure 3 presents on the left side the biggest difference in favour of AutoML and the right side in favour of a statistical model.

## 5. Conclusion and future research agenda

The results show a mixed picture. On the one hand, some AutoML tools perform well in certain scenarios and far better than statistical models in a few cases. However, in other scenarios AutoML is less accurate than statistical models, even in certain scenarios by far. AutoML, namely the tools by Dataiku and Microsoft Azure, tend to predict more stable in short-term predictions, but ARIMA has a slightly lower median of MASE than Dataiku across all product groups. Especially for case 1, AutoML achieves for six of seven product groups the lowest MASE value. This could be due to the fact that descriptive features are part of the dataset. For case 2, where only additional time information exists, only in three out of ten product groups, an AutoML tool was prior to the rest. However, in average all models are performing worse than in case 1. The results of the long-term predictions for case 1 show that at least one AutoML tool performs best for four groups and that a statistical model is the preferred choice for the other three groups. The assumed advantage through more descriptive feature is not as clear as for the short-term prediction horizon. In case 2, at least one AutoML model predicts the demand of most product groups (6 out of 10) more precisely than a statistical model. However, the results of ARIMA and exponential smoothing deviate not as much as of the AutoML tools. The AutoML tool of Google seems to be most sensitive to the balance of data as some product groups were more frequently demanded than others. Especially in case 2, prediction accuracies for less demanded product groups (e.g. 001, 011, 015, 021, 024) are far above a MASE value of 1. Nevertheless, Google's tool also trains the most accurate model with a MASE of 0.127 (product group Food 3). As this paper shows, AutoML can support on preparation tasks, modelling and the associated optimization of hyper-parameters as well as the evaluation of models. However, as this analysis expresses prediction results are not fully reliable yet. Further investigations should take place to understand the differences in performance. As a ML pipeline includes several assumptions, some aspects for further investigations are outlined: Firstly, modelling with more features should be conducted, to test the hypothesis that AutoML is in favour when predicting on a multivariate basis. The herein analysed datasets have only few features, which presumably explains why the models partially reach their limits when it comes to long-term forecasting. In addition, the test and training split should be varied to get further insights into an eventual over- or underfitting of a ML model. Also, to understand the sensitivity of AutoML tools further demand patterns, longer history of data, different settings of data preparation (e.g. keeping missing values), single and global models, prediction horizons and different case studies should be investigated. Moreover, further statistical models (e.g. ARIMAX), manually trained ML-models and fuzzy models could function as additional benchmarks that should be evaluated on prediction accuracy, running and implementation time. In summary, the results can help to find a sweet spot for the use of AutoML. This howsoever highly depends on the manufacturing setup, i.e. the required prediction horizon results e.g. from procurement time for materials, storage capacity and production lead times. It should be noted that the analyses are repeated regularly to examine the progress of AutoML/ ML over time.

Thus, in conclusion, the research question whether AutoML can support demand planning in PPC can be answered as following: this paper shows especially for short-term predictions good results for AutoML. However, for some demand patterns and less demanded product groups, the accuracy was not sufficient. Thus, AutoML can function for prototyping and can be part in business processes. When implementing into business processes the chosen AutoML tool should be regularly tested against a benchmark of different ML and statistical models. AutoML can significantly reduce the time spend to analyse the provided data as well as train and optimize different algorithms and can therefore be a first step for companies to test ML in their business environment. Therefore, it can help to ensure that domain expertise is effectively reflected in data-driven models by enabling domain experts to use ML without having extensive data science knowledge. Nevertheless, ML and AutoML comes at the cost of models that are more complex and use more computing power in comparison to statistical models [32]. With the use of AutoML tools the understanding of the 'engine' behind the models, for instance how the feature engineering or optimizing of the parameters from the algorithm take place, becomes less transparent as they are for most tools not visible in the user interface. In the future, aspects such as user-friendliness, transparency and trustworthiness of workflows should be considered next to the tools' prediction accuracy, running and implementation time as well as robustness.

## Acknowledgements

Funded by the Lower Saxony Ministry of Science and Culture under grant number ZN3489 within the Lower Saxony "Vorab" of the Volkswagen Foundation and supported by the Center for Digital Innovations (ZDIN).

## References

[1] Gentsch, P., 2019. Künstliche Intelligenz für Sales, Marketing und Service, 2nd ed. Springer Fachmedien Wiesbaden, Wiesbaden.

[2] Chui, M., Manyika, J., Miremadi, M., Henke, N., Chung, R., Nel, P., Malhotra, S., 2018. Notes from the AI frontier: Insights from hundreds of use-cases, 36 pp. https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-applications-and-value-of-deep-learning. Accessed 16 July 2021.

[3] BMWi, 2015. Erschließen der Potenziale der Anwendung von "Industrie 4.0" im Mittelstand. Bundesministerium für Wirtschaft und Energie. https://www.bmwi.de/Redaktion/DE/Publikationen/Studien/erschliessen-der-potenziale-der-anwendung-von-industrie-4-0-im-mittelstand.html. Accessed 25 August 2021.

[4] Elshawi, R., Sakr, S., 2020. Automated Machine Learning: Techniques and Frameworks, in: Kutsche, R.-D., Zimányi, E. (Eds.), Big Data Management and Analytics. 9th European Summer School, eBISS 2019, Berlin, Germany, June 30 – July 5, 2019, Revised Selected Papers, vol. 390, 1st ed. 2020 ed. Springer International Publishing, Cham, pp. 40–69.

[5] He, X., Zhao, K., Chu, X., 2020. AutoML: A survey of the state-of-the-art. Knowledge-Based Systems.

[6] Deloitte. State of AI in the Enterprise - 3rd Edition. https://www2.deloitte.com/de/de/pages/technology-media-and-telecommunications/articles/ki-studie-2020.html. Accessed 17 January 2022.

[7] PwC. Künstliche Intelligenz in Unternehmen. https://www.pwc.de/de/digitale-transformation/kuenstliche-intelligenz/kuenstliche-intelligenz-in-unternehmen.html. Accessed 17 January 2022.

[8] Hutter, F., Kotthoff, L., Vanschoren, J., 2019. Automated Machine Learning. Springer Intl Publishing, Cham.

[9] Bender, J., Ovtcharova, J., 2021. Prototyping Machine-Learning-Supported Lead Time Prediction Using AutoML. Procedia Computer Science 180, 649–655.

[10] Waring, J., Lindvall, C., Umeton, R., 2020. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artificial intelligence in medicine (104), 1–12.

[11] Dai, Y., Huang, J., 2021. A Sales Prediction Method Based on LSTM with Hyper-Parameter Search. J. Phys.: Conf. Ser. (1756).

[12] Ford, J., Nava, C., Tan, J., Sadler, B., 2020. Automated Machine Learning Framework for Demand Forecasting in Wholesale Beverage Alcohol Distribution. SMU Data Science Review (3).

[13] Henzel, J., Sikora, M., 2020. Gradient Boosting and Deep Learning Models Approach to Forecasting Promotions Efficiency in FMCG Retail, in: Rutkowski, L., Scherer, R., Korytkowski, M., Pedrycz, W., Tadeusiewicz, R., Zurada, J.M. (Eds.), Artificial Intelligence and Soft Computing, vol. 12416. Springer International Publishing, Cham, pp. 336–345.

[14] Gonçalves, J.N., Cortez, P., Carvalho, M.S., Frazão, N.M., 2020. A multivariate approach for multi-step demand forecasting in assembly industries: Empirical evidence from an automotive supply chain. Decision Support Systems (142).

[15] Hutter, F., Caruana, R., Bardenet, R., Bilenko, M., Guyon, I., Kegl, B., Larochelle H., 2014. AutoML 2014 workshop. https://sites.google.com/site/automlwsicml14/. Accessed 29 November 2021.

[16] Zöller, M.-A., Huber, M.F., 2021. Benchmark and Survey of Automated Machine Learning Frameworks. Journal of Artificial Intelligence Research (70), 409–472.

[17] Yao, Q., Wang, M., Chen, Y., Dai, W., Li, Y.-F., Tu, W.-W., Yang, Q., Yu, Y., 2018. Taking Human out of Learning Applications: A Survey on Automated Machine Learning, 20 pp. Accessed 29 November 2021.

[18] Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer, 2000. CRISP-DM 1.0 Step-by-step data mining guide.

[19] Alegado, R.T., Tumibay, G.M., 2020. Statistical and Machine Learning Methods for Vaccine Demand Forecasting: A Comparative Analysis. Journal of Computer and Communications (Vol.8), 37–49.

[20] Maaß, D., Spruit, M., Waal, P. de, 2014. Improving short-term demand forecasting for short-lifecycle consumer products with data mining techniques. Decis. Anal. 1 (1).

[21] Schreiber, L., Moroff, N., 2020. Machine Learning versus Statistical Methods in Demand Planning for Energy-Efficient Supply Chains. ICoMS 2020: Proceedings of the 2020 3rd International Conference on Mathematics and Statistics, 17–23.

[22] Witten, I.H., Frank, E., Hall, M.A., 2011. Data mining: Practical machine learning tools and techniques, Third edition ed. Morgan Kaufmann, Amsterdam.

[23] Cerqueira, V., Torgo, L., Soares, C., 2019. Machine Learning vs Statistical Methods for Time Series Forecasting: Size Matters, 9 pp. http://arxiv.org/pdf/1909.13316v1. Accessed 25 May 2021.

[24] Hyndman, R.J., Athanasopoulos, G., 2021. Forecasting: principles and practice. OTexts.com/fpp3. Accessed 2 December 2021.

[25] Wiendahl, H.-P., Wiendahl, H.-H., 2019. Betriebsorganisation für Ingenieure, 9., vollständig überarbeitete Auflage ed. Hanser, München.

[26] FelixZhao, 2017. Forecasts for Product Demand: Make Accurate Forecasts for Thousands of Different Products. Data. https://www.kaggle.com/felixzhao/productdemandforecasting. Accessed 2 December 2021.

[27] University of Nicosia, 2020. M5 Forecasting - Accuracy: Estimate the unit sales of Walmart retail goods. data. https://www.kaggle.com/c/m5-forecasting-accuracy/data. Accessed 2 December 2021.

[28] Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer New York, New York, NY.

[29] Javeri, I.Y., Toutiaee, M., Arpinar, I.B., Miller, T.W., Miller, J.A., 2021. Improving Neural Networks for Time Series Forecasting using Data Augmentation and AutoML. http://arxiv.org/pdf/2103.01992v3. Accessed 30 June 2021.

[30] Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. International Journal of Forecasting 22 (4), 679–688.

[31] Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2018. Statistical and Machine Learning forecasting methods: Concerns and ways forward. PloS one 13 (3).

[32] Russell, S.J., Norvig, P., 2016. Artificial intelligence: A modern approach, Third edition, Global edition ed. Pearson, Boston, Columbus, Indianapolis.

## Biography

**Kathrin Julia Kramer** (*1991) is an Academic Assistant and PhD-student in the field of production management at the Institute of Product and Process Innovation (PPI) at the Leuphana University Lueneburg since 2019. Her research focus is on the practical application of machine learning methods in production planning and control.

**Niclas Behn** (*1993) is a Graduate Trainee in Operations. In 2021, he completed his master degree in Management & Engineering at the Institute of Product and Process Innovation (PPI) at the Leuphana University Lueneburg.

**Matthias Schmidt** (*1978) studied industrial engineering at the Leibniz University Hannover and subsequently worked as a research associate at the Institute of Production Systems and Logistics (IFA). After completing his doctorate in engineering, he became head of Research and Industry of the IFA and received his habilitation. Since 2018, he holds the chair of production management at the Institute for Product and Process Innovation (PPI) at the Leuphana University of Lueneburg. In addition, he became the head of the PPI in 2019.