Data Science Meets High-Tech Manufacturing – The BTW 2021 Data Science Challenge

Lucas Woltmann¹ · Peter Volk² · Michael Dinzinger³ · Lukas Gräff³ · Sebastian Strasser³ · Johannes Schildgen³ · Claudio Hartmann¹ · Wolfgang Lehner¹

Received: 15 October 2021 / Accepted: 18 November 2021 / Published online: 21 December 2021
© The Author(s) 2021

Abstract

For its third installment, the Data Science Challenge of the 19th symposium “Database Systems for Business, Technology and Web” (BTW) of the Gesellschaft für Informatik (GI) tackled the problem of predictive energy management in large production facilities. For the first time, this year’s challenge was organized as a cooperation between Technische Universität Dresden, GlobalFoundries, and ScaDS.AI Dresden/Leipzig. The Challenge’s participants were given real-world production and energy data from the semiconductor manufacturer GlobalFoundries and had to solve the problem of predicting the energy consumption for production equipment. The usage of real-world data gave the participants a hands-on experience of challenges in Big Data integration and analysis. After a leaderboard-based preselection round, the accepted participants presented their approach to an expert jury and audience in a hybrid format. In this article, we give an overview of the main points of the Data Science Challenge, like organization and problem description. Additionally, the winning team presents its solution.

Keywords Data Science · BTW 2021

1 Introduction

For the third time, the Data Science Challenge (DSC) took place as a part of the conference *Datenbanksysteme für Business, Technologie und Web 2021* (BTW 2021)¹. The main idea behind this challenge is that participants are able to test their data science knowledge on a real-world data set and tackle problems with practical aspects. As in previous years [6, 7], the BTW organizers consider the DSC as an attractive addition to the traditional conference formats.

In the context of the DSC, a data set and a task are provided to the public. The task needs to be solved by the participants with information concerning the data set and techniques related to data science and machine learning (ML). For this, the participants have no limitations on what to use or how to tackle the problem. In previous DSC events, the participants used a diverse collection of approaches to model the usage of bike-sharing services in New York [7] or to predict the suspended particulate matter in German cities [6]. This shows how manifold the problems and solutions of a DSC can be.

✉ Lucas Woltmann
lucas.woltmann@tu-dresden.de

Peter Volk
Peter.Volk@gf.com

Michael Dinzinger
michael.dinzinger@st.oth-regensburg.de

Lukas Gräff
lukas-graef@st.oth-regensburg.de

Sebastian Strasser
sebastian.strasser@st.oth-regensburg.de

Johannes Schildgen
johannes.schildgen@oth-regensburg.de

Claudio Hartmann
claudio.hartmann@tu-dresden.de

Wolfgang Lehner
wolfgang.lehner@tu-dresden.de

¹ Dresden Database Research Group, Technische Universität Dresden, Dresden, Germany

² GlobalFoundries Dresden GmbH, Dresden, Germany

³ Fakultät für Informatik und Mathematik, OTH Regensburg, Regensburg, Germany

¹ <https://sites.google.com/view/btw-2021-tud/>.

In this year's competition, the encompassing aspect of the *Energiewende* and ecological energy consumption was at the center of attention. With the ever-growing demand of large industries to save and preserve energy in their production facilities, the need for a smart solution to predict energy consumption is very high. This includes better energy management, motivated by the *Energiewende* and regulated by the ISO 50001 [5], to make energy consumption more efficient and to prepare companies for upcoming challenges within the energy market. To support the companies within this field, the DSC aimed to solve the problem of predicting energy consumption at the tool level for a large semiconductor plant. The data set was kindly provided by GlobalFoundries², one of the world's largest manufacturers of semiconductors. With production data (also known as recipes) and energy data, the participants were asked to predict the energy consumption for different tools in the GlobalFoundries' plant in Dresden.

The contest was organized as a cooperation between GlobalFoundries, the Database Systems Group at TU Dresden, and ScaDS.AI Dresden/Leipzig. During the first phase, from January to July 2021, the participants' solutions were scored with a quantitative measure on a leader board. The best two teams were moved to the second round where they had to defend their ideas and solutions in front of a jury of scientists and energy domain experts. The winning team and the runner-up were awarded prizes totaling 1000 €.

The remainder of this paper is structured as follows: In the next section, we provide an overview of the actual task and the data set. After that, the winning team presents its solution. Finally, we close with some final thoughts and acknowledgments.

2 Task Description

With the high energy demand in a semiconductor production facility (short: fab), the plant operator is required to measure, monitor, and control the energy consumption of the production tools within the fab. This whole process is commonly known as energy management, sometimes referred to as ISO 50001. GlobalFoundries has a comprehensive energy management system, including the measurements of several tools and tool groups. However, not every tool can be monitored, so only representative tools are measured for their consumption. Therefore, the main task of the DSC was to build a model (or a collection of models) to predict the energy consumption for a collection of tools based on former energy measurements and work-in-progress (WIP) data. The WIP data contains useful information because the process running on the tool directly

Fig. 1 An overview of the conceptual problem description

impacts its energy consumption. Fig. 1 visualizes the problem description. It shows two tool groups (blue and yellow) with similar properties in two production areas. Only some tools of each group are monitored with energy meters W , but the task requires that a model f should work for any tool with the same type it was trained on.

From this task, several challenges arise. First, the energy and WIP data are in separate data sets and need to be joined using timestamps. However, the timestamps are present having various granularities that require a complex join logic. Second, there are parallel processes on a single tool, each having a different impact on the energy consumption. The contestants need to find a solution to differentiate between parallel WIP processes. Third, there are different tools and tool types in training and test data. Given that not all tools can be monitored, a solution should be adaptable to new tools within a tool group. These challenges are specific to this particular problem. Of course, the participants also have to solve the typical ML challenges, like training performance and hyperparameter tuning, that arise in any Data Science task.

The aforementioned challenges should be tackled using two data sets: energy consumption and WIP data. The information in both data sets is obfuscated to eliminate the possibility to gain any information about the internal processes of GlobalFoundries. The energy data set contains five-second measurement intervals containing minimum, maximum, and average energy in Watt-hours. All three measurements need to be predicted. The training data contains 100 million and the test data 26 million measurements. The WIP data is a representation of the recipe currently in production on a tool. These recipes usually have a duration of hours or days. They include information about the start and the end of the process and its characteristics. Here, we provide 7 million training and 3 million test tuples.

² <https://gfdresden.de/>.

3 Winning Solution

The following part introduces and discusses three strategies for forecasting the energy consumption of tools used in semiconductor manufacturing. These strategies are applied and evaluated on the aforementioned production data from GlobalFoundries. By grouping tools with similar properties, visual and explorative analysis led to the discovery of patterns within the energy consumption. In particular, one of these groups contains patterns that coincided with a peak in energy consumption with a high probability. This observation led to a forecasting method with good results. Additionally, methods to forecast tools with no clear patterns are also covered.

3.1 The Assignment and its Challenges

At the start of the challenge, two CSV-formatted files containing real but obscured data were supplied by GlobalFoundries. The first file contains the energy consumption of 17 tools, measured in intervals of five seconds over a time span of one year. This results in approximately 100 million records. For each five-second interval, the minimum, average, and maximum energy consumption during the corresponding time period are given. The last quarter of the data set is filled with NULL values, thus representing the time frame to predict. The second file consists of approximately 10 million records and contains WIP processing information of workpieces during the relevant time span. In the context of semiconductor manufacturing, a workpiece mostly corresponds to a wafer.

With both files containing about 110 million records in total, the first obvious challenge is the amount of data. This circumstance has multiple consequences. First, performing analysis directly on the CSV files (e.g., using Python) is not reasonable. Therefore, a suitable way of storing and accessing the records is necessary. Second, the amount of data also influences the frequency of the forecasting horizon: For each tool, an energy value is measured every 5 seconds, resulting in about 1.5 million time steps over 3 months for each tool, which have to be predicted.

Another challenge is to combine the processing information found in the WIP file with the corresponding energy measurements. This is caused by several processes running in parallel on each tool. Consequently, each measurement can be assigned to multiple entries in the WIP data.

3.2 Data Preparation and Management

Data preparation started with loading the CSV files into a relational database which facilitates further data processing. The two resulting tables are named ENERGY and WIP, in accordance to the contents of the two CSV files. As the initially provided data contains erroneous rows and inconsistencies, it was necessary to perform some kind of data cleaning using SQL commands. One major improvement in data quality was achieved by aligning the time zones of both tables. Furthermore, time inconsistencies resulting from the switch to daylight-saving time were recognized and cleaned up. Another result was the detection and removal of outliers. Specifically, processes lasting multiple weeks were detected and removed as the changes in the energy profile are all of very short duration. Therefore, such long-lasting processes seem to have an insignificant impact on the energy consumption and, as a consequence, can be omitted in the upcoming forecasting task. As a result of these preparation measures, data consistency was improved.

To connect the process information with the measurements of the energy meters, a temporal join between the tables WIP and ENERGY was performed. As multiple processes run on multiple sub-components on the same equipment in parallel, a join containing all processes running in a five-second interval would produce billions of entries. Hence, it was necessary to aggregate the data without creating even more entries. Specifically, the approach taken was based on determining all processes running on a given tool for a given timestamp. This was achieved by counting the occurrences of processes within the five-second time frame. These counts are then added as additional columns to the energy table.

Fig. 2 Energy profile of tool 4218 including the prediction interval. The differently colored peaks correspond to the minimum and maximum energy consumption

3.3 Methods

In order to reduce the complexity of the forecasting task, the team divided the set of different tools into separate categories. They were split into groups with similar properties, by comparing, among other things, their energy profiles and the set of entities they consist of. The visualization of the energy consumption proved to be instrumental in gaining insights and recognizing patterns within the data. Fig. 2 shows the complete energy profile of the tool with ID 4218 with the last quarter of the data set being the time frame to forecast.

For predicting the data, multiple forecasting strategies were utilized: An approach based on finding manual patterns, a strategy based on regression, and a naive approach.

The most successful approach, *manual pattern finding*, is based on the observation that the energy profiles of several tools consist of a baseline and a varying number of peaks. These peaks can be separated into different categories, whereby each kind of peak is indicated by certain events in the WIP data. These observations can be used to forecast the energy measurements: The WIP data of each tool is scanned for the events which are known to cause a peak. Whenever such an event occurs, the gradient of energy consumption which is typical for this type of event is inserted. After the peak insertion process, all remaining values are filled with a suitable constant. The main task is therefore the search for patterns in the WIP data that cause a peak in the energy consumption. As the name of this

strategy reveals, the patterns have to be found manually as this specific challenge is not covered by any off-the-shelf machine learning algorithm.

The described procedure is illustrated in Fig. 3. It performs well on tools with a predictable energy profile, e.g. on the equipment with ID 4218: The prediction yielded accurate results with a SMAPE (Symmetric mean absolute percentage error) of approximately 3.5% on all three target values. It demonstrates that this approach delivers quite accurate results and, consequently, convinces due to a reasonably good SMAPE.

Another advantage of this proceeding is its simple implementation. As there is no join required, this method is able to deal with the high complexity and quantity of the underlying data. The drawback of this approach is, however, that the patterns in the WIP data, which are causing the energy peaks, have to be discovered first. Fig. 4 illustrates the relevant time periods for the peak recognition as the red spans. Ideally, for every kind of peak, a representative event can be located in the WIP data. For example, an event can be a process on a certain entity. Unlike most advanced machine learning models, in this approach, the pattern recognition is separated from the prediction and is not automated. However, by creating supportive scripts, the peak recognition was done in a semi-automated way.

Some tools do not follow the strict pattern that was assumed in the approach mentioned prior. To forecast the values of these tools, *regression models* (e.g., *Random Forests* [1] and *Huber Regression* [3]) were used. For each

Fig. 3 Manual Pattern Finding: The continuous plot in the background is the original energy profile. The peaks are modeled with peak insertion and parts between the peaks with median prediction

Fig. 4 Peak Finding – Dots represent peaks

Fig. 5 Forecast energy values over the span of a month using Random Forests. The forecast values overlay the real values

Table 1 Overview of the strategies and the tools they are applied on

Naive Approach	1463, 1464, 4207, 4213, 4550
Regression	4103, 4104, 4113, 4119
Pattern Finding	4214, 4215, 4216, 4217, 4218, 4219, 4432, 4433, 4531

tool, three models were trained with either the minimum, average, or maximum value as the target. The features used in the training phase were mostly extracted from the joined table that is described in Sect. 3.2. The most important features are the counts of activities executed and the number of wafers processed on the equipment in one time frame. An illustration of the regression-based forecasting can be found in Fig. 5.

In the *naive approach*, the median of the energy measurements over the whole time period is used. This prediction was only applied in special scenarios, e.g., when no process data was available. In these cases, the median performed better than any other approach.

Table 1 shows which of the three presented strategies is used for the given set of tools. The strategies were primarily selected in order to achieve the best possible SMAPE. Whereas the naive approach and regression could be used for predicting the energy consumption of all tools, pattern finding is only possible in case there is a known correlation between events in the WIP data of a tool and peaks in its energy profile. In return, the pattern recognition delivers the best prediction results using these tools.

Alternative approaches such as autoregressive models [4] and neural networks [2] were considered and tested, but proved to be difficult to apply in this scenario, mainly due to the high volume of the data.

3.4 Tools and Frameworks

The main tools used in this challenge were the relational database Exasol, for general data management, and the programming language Python with its extensive libraries and frameworks. The database was hosted on a dedicated server with 88 GB RAM. Python was used for two tasks: On one

hand to perform exploratory analysis using Pandas³ and Matplotlib⁴. On the other hand, Python was used to perform the forecasting using Pandas, Numpy⁵, ScikitLearn⁶ as well as sktime⁷.

3.5 Summary

The main challenge exhibited by this competition was the sheer amount of supplied data. As such, this had multiple consequences for approaching this task. First of all, it was not trivial to combine the two tables using temporal joins and therefore, every SQL command had to be planned carefully beforehand. Secondly, the huge forecasting horizon of the energy table meant that classical approaches for time-series prediction were not appropriate to solve this challenge. Finally, the amount of data meant that simple methods were preferable, as they ensured a smaller run time compared to complex models, e.g., neural networks. In summary, the best results were achieved using simple hypotheses to explain the occurrence of peaks.

4 Final Words

The DSC has become a well-established part of BTW conference covering many interesting topics in applied Data Science. This year's challenge shows that there is a high demand for industry players for solving a wide variety of data-driven problems with crowd-sourcing. The topic of energy management is highly aligned with real-world problems, therefore both the industry partner and the participants can profit from the implemented solutions. The significance of the optimal use of energy will grow in importance over the next years, so we expect more work and cooperation especially in the area of Data Science. We are looking forward

³ <https://pandas.pydata.org/>.

⁴ <https://matplotlib.org/>.

⁵ <https://numpy.org/>.

⁶ <https://scikit-learn.org/>.

⁷ <https://www.sktime.org/>.

to new challenging tasks during the next DSC at the BTW 2023 in Dresden.

Last but not least, we would like to thank Holger Meyer (Universität Rostock), Grit Herrmann (GlobalFoundries), Markus Arend (Energiemanagement Dr. Markus Arend), Ulrike Schöbel (Technische Universität Dresden), and Corina Weissbach (ScaDS.AI Dresden/Leipzig) for their commitment and support during the DSC.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
2. Géron A (2019) Hands-on machine learning with Scikit-learn and tensorflow: concepts, tools, and techniques to build intelligent systems, 2nd edn. O'Reilly Media,
3. Huber PJ (1973) Robust regression: asymptotics, conjectures and Monte Carlo. *Ann Stat* 1(5):799–821
4. Hyndman R, Athanasopoulos G (2018) Forecasting: principles and practice, 2nd edn. OTexts,
5. ISO Central Secretary (2018) ISO 50001:2018 – Energy management systems. Standard, International Organization for Standardization, Geneva, CH
6. Meyer HJ, Grunert H, Waizenegger T et al (2019) Particulate matter matters – the data science challenge @ BTW 2019. *Datenbank Spektrum* 19(3):165–182. <https://doi.org/10.1007/s13222-019-00322-x>
7. Waizenegger T (2017) BTW 2017 data science challenge (SDSC17). In: Mitschang B, Ritter N, Schwarz H et al (eds) *Datenbanksysteme für Business, Technologie und Web (BTW 2017)* 17. Fachtagung des GI-Fachbereichs “Datenbanken und Informationssysteme” (DBIS), Stuttgart, 6.–10. März 2017. Workshopband, LNI, vol P-266. GI, Germany, pp 405–406. <https://dl.gi.de/20.500.12116/938>