Chemical Laboratories 4.0: A Two-Stage Machine Learning System for Predicting the Arrival of Samples

António João Silva, Paulo Cortez, André Pilastri

Chemical Laboratories 4.0: A Two-stage Machine Learning System for Predicting the Arrival of Samples

António João Silva[1,2][0000−0003−4747−5073], Paulo Cortez[2][0000−0002−7991−2090], and André Pilastri[1][0000−0002−4380−3220]

[1] Centro de Computação Gráfica, 4804-533 Guimarães, Portugal
{antonio.silva,andre.pilastri}@ccg.pt
[2] ALGORITMI Center, Department of Information Systems, University of Minho, 4804-533 Guimarães, Portugal
pcortez@dsi.uminho.pt

Abstract. This paper presents a two-stage Machine Learning (ML) model to predict the arrival time of In-Process Control (IPC) samples at the quality testing laboratories of a chemical company. The model was developed using three iterations of the CRoss-Industry Standard Process for Data Mining (CRISP-DM) methodology, each focusing on a different regression approach. To reduce the ML analyst effort, an Automated Machine Learning (AutoML) was adopted during the modeling stage of CRISP-DM. The AutoML was set to select the best among six distinct state-of-the-art regression algorithms. Using recent real-world data, the three main regression approaches were compared, showing that the proposed two-stage ML model is competitive and provides interesting predictions to support the laboratory management decisions (e.g., preparation of testing instruments). In particular, the proposed method can accurately predict 70% of the examples under a tolerance of 4 time units.

Keywords: Automated Machine Learning · Industry 4.0 · Regression.

1 Introduction

The Industry 4.0 concept assumes a high usage of Artificial Intelligence (AI), where industrial physical processes generate data that can be analyzed by Business Analytics, namely Data Mining (DM) and Machine Learning (ML) techniques, aiming to improve the factory efficiency (e.g., reduce costs, enhance production levels) [21]. This concept is transforming the Chemical industry, which has a large impact in the world economy (e.g., petrochemicals, pharmaceuticals).

In this work, we address a relevant Business Analytics need of a chemical company, which is adopting a Industry 4.0 transformation. To ensure the quality of the products being manufactured, samples taken from the company production processes need to be tested in laboratories. The tests assure that the

products are compliant with quality standards, allowing their usage by the company clients. Under this context, predicting the arrival of production samples at the laboratory is a key issue, since it helps in the allocation of equipment and human resources. Aiming to solve this task, this paper presents a novel two-stage ML prediction system, which was developed during the implementation of a CRoss-Industry Standard Process for DM (CRISP-DM) [25] project that included three iterations, each focusing on a distinct regression strategy. During the modeling stage of the three CRISP-DM iterations, an Automated ML (AutoML) [12] procedure was adopted, allowing to compare and configure six state-of-the-art ML algorithms.

The paper is structured as follows. Section 2 describes the related work. The business task, data and proposed approach are presented in Section 3. The obtained results are shown in Section 4. Finally, Section 5 concludes the paper.

2 Related Work

In recent years, there has been an increased interest in the field of AI, due to the rise of data, computational power and sophisticated learning algorithms (e.g., Deep Learning) [9]. Following the Industry 4.0 revolution [21], many factories now are generating data that can be analyzed by DM and ML techniques in order to support managerial decision-making. Yet, several real-world DM projects tend to fail due to a misalignment between business needs and ML analyses [10]. The CRISP-DM is an open standard and robust methodology that was specifically developed to reduce this misalignment and increase the success of DM projects [25]. CRISP-DM includes six stages that are executed through several iterations and that involve both business and ML experts: business understanding, data understanding, data preparation, modeling, evaluation and deployment. CRISP-DM is a popular methodology. For instance, it has been applied to the Banking [18] and Health Care [3] domains.

Regarding the analyzed chemical industry, the quality testing laboratories are mostly managed manually, with the usage of Information Technology (IT) being more focused on storing the test values rather than the process [22, 16]. Moreover, the data is typically spread through different databases what work as information silos (e.g., production, laboratory testing), thus it is difficult to have an easy access to all data under a single version of the truth. By adopting the Industry 4.0 concept, which assumes a better usage of IT, there is a potential gain to optimize the management of the chemical laboratories. In this work, we describe one aspect of the Industry 4.0 transformation that is being conducted by a chemical company. It corresponds to the result of implementing a CRISP-DM project that uses both production and laboratory testing databases.

In terms of Predictive Analytics applied to the industry, most studies target predictive maintenance via several ML algorithms, such as Random Forest (RF) [4], Neural Networks (NN) [23] and Gradient Boosting Machines (GBM) [17]. There are also studies about non maintenance prediction applications, such as: the classification of quality products produced by injection molding processes

via Boosting, RF and NN models [5]; and estimation of endpoint temperature and chemical concentration of a furnace when producing low-carbon steel using RF and ridge regression algorithms [19]. All these studies require the selection and configuration of the right ML algorithm, which often depends on the ML expert knowledge and that involves the usage of heuristics or trial-and-error experiments [14]. In order to avoid this time-consuming procedure (in terms of the ML expert effort), we adopt an AutoML [12] during the modeling stage of the CRISP-DM. This systematization and automation the ML model selection provides two main advantages. First, it alleviates the effort of the ML analyst, allowing to focus on other ML aspects in order to provide a better business value. In particular, in this paper, it allowed to implement more iterations of the CRISP-DM methodology, which was helpful to design the proposed two-stage ML model. Second, it reduces the ML maintenance effort, since the ML can be retrained automatically, as new data arrives, which is advantageous for the analyzed company.

3 Materials and Methods

3.1 Business Task

The analyzed chemical company produces several products, in batches. During the production-batch execution process, a sequence of samples, called In-Process Control (IPC), are selected for quality laboratory inspection, in order to ensure that the production process is running as expected. In terms of the chemical laboratories, the IPC samples have the highest priority, because the production process can not continue without their approval. A fixed amount of IPC samples are selected from each production-batch ($s \in \{1, ..., IPC_{\max}\}$). The production information system registers several attributes related to the IPC sample production, including its initial production time, denoted here as IPC production time PT_s. One by one, the IPC samples arrive at the laboratory at time LT_s, under irregular intervals that are difficult to be estimated in advance.

The business goal is thus the non-trivial task of predicting of arrival time for each IPC sample at the chemical laboratories. Solving this task efficiently allows a better management of the laboratory equipment and human resources. For instance, some IPC quality tests require a setup time, in which the analysts need to prepare in advance the laboratory testing instruments. The business goal was addressed as a regression task, under two main target goals. In the first CRISP-DM iteration, we only used laboratory temporal data and the target goal was defined as predict $y_1 = LT_{s+1} - LTs$, which corresponds to the time lag between the next IPC sample arrival (LT_{s+1}) and the current (already known) laboratory sample arrival (LT_s). In the second and third CRISP-DM iterations, we explored production temporal data, predicting $y_2 = LT_s - PT_s$, where the laboratory arrival time can be immediately estimated once the IPC sample starts its production.

3.2 Data Understanding and Preparation

We used an Extract, Transform, load (ETL) procedure to merge the relevant data from two main databases related with the production and laboratory testing information systems, populating an integrated and business oriented data warehouse system. The ETL resulted in a raw file with 226,929 rows and 33 columns regarding all laboratory samples that were analyzed during a three-year time period. The data warehouse was further filtered in order to contain rows related with IPC samples and with complete values in terms of the input and output attributes (Table 1), leading to a dataset with 26,611 instances. The input variables were manually selected and defined from the filtered raw file using expert domain knowledge, obtained by interacting with the chemistry experts. Due the complexity of the chemical factory processes and information system integration issues, it was not possible to have access to a more richer set of data features (e.g., which components and machines were used to produce the samples). Thus, the resulting set of 8 inputs is rather small, which makes more challenging the prediction task. Both output targets were computed using a particular time unit, which is not disclosed here due to business privacy issues.

Table 1. Summary of the data attributes.

Input Attributes:		
Name	**Description**	**Range**
day	day of the week when the production-batch started	{1,...,7}
month	month when the production-batch started	{1,...,12}
product	product type (nominal code)	155 levels
version	version of the product (numeric)	{1,...,108}
grade	product grade (nominal, related with the lab tests)	15 levels
stage	product stage (nominal, related with the lab tests)	1,272 levels
batch	batch identification of the product (nominal)	925 levels
s	sequence number of the sample ($s \in \{1, ..., IPC_{\max}\}$)	{1,...,169}
Output Targets:		
Name	**Description**	**Range**
y_1	time lag arrival of two consecutive samples	[0.2,5315.3]
y_2	time lag between PT_s and LT_s	[0.0,3270.0]

3.3 Machine Learning Models

In terms of computational environment, we adopted the R tool and its `rminer` package [8] for data manipulation and ML result evaluation, while the AutoML adopts the H2O implementation [7]. The AutoML procedure was configured to select the regression model and its hyperparameters based on the best Root Mean Squared Error (RMSE) computed using a validation set that is obtained by

applying an internal 10-fold cross-validation method over the training data. All computational experiments were executed on the same personal computer and each individual ML model was trained up to a maximum running time of 3,600 seconds. Once a ML model is selected, the model was retrained with all training data. As in [11], the AutoML was configured to include a total of 6 distinct regression algorithms: RF, Extremely Randomized Trees (XRT), Generalized Linear Model (GLM), GBM, XGBoost (XG) and a Stacked Ensemble (SE). The RF is a popular ensemble method that combines a large number of decision trees based on bagging and random selection of input features [15]. The XRT algorithm extends the RF approach by randomly selecting the decision thresholds of the tree nodes [13]. GLM estimates regression models for exponential distributions (e.g., Gaussian, Poisson, gamma) [15]. The GBM algorithm is a based on a generalization of tree boosting, sequentially building regression trees for all data features [15]. XG is another ensemble tree method that uses boosting to enhance the prediction results [6]. The SE method, also known as stacked regression [2], combines the predictions of different base learners by using a second-level ML algorithm. The H2O implementation [7] uses the following AutoML setup: RF and XRT – set with the default hyperparameters; GLM - grid search used to set one hyperparameter (*alpha*, a regularization parameter); GBM and XG – grid search used to tune nine and ten hyperparameters (e.g., number of trees, maximum depth, minimum rows); SE – all five algorithms (RF, XRT, GLM, GBM, XG) are used as base learners and the individual predictions are weighted by using a second-level GLM learner. For the ML algorithms that require numeric inputs (e.g., GLM), the nominal inputs (e.g., product, grade) are previously transformed by using the standard one-hot encoding, which assigns one boolean input per categorical level. For instance, a categorical feature with three levels ($\{a,b,c\}$) is encoded as: a =(1,0,0), b =(0,1,0) and c =(0,0,1).

A total of three CRISP-DM iterations were executed, aiming to improve the regression results and the potential value of the ML models. The first CRISP-DM iteration targeted the y_1 output, while the second and third CRISP-DM iterations approached y_2, under two variants. The y_1 target is assumes that at least one IPC sample from the production-batch as arrived at the laboratory. The trained ML model can be used each time new sample arrives, allowing to estimate when the next sample will be delivered ($\widehat{y}_1$). A different perspective is adopted by the y_2 target, since the fitted ML model can be applied to predict the laboratory sample arrival once an IPC sample production has started. The model employed in the second CRISP-DM iteration uses a simple regression with a single ML model ($\widehat{y}_2$). During the evaluation stage of the second CRISP-DM iteration, we identified that there were some high prediction errors, in particular when predicting the arrival times for the first sample of the production-batch ($s = 1$). In order to check if we could improve these results, a third CRISP-DM iteration was executed, in which we specialize two distinct ML models (α and β). The first ML model (α) is trained using only the first product-batch sample examples ($s = 1$) and thus the fitted model includes only seven input attributes ({day, month, product, version, grade, stage,batch}). The second model (β) is

only activated when producing the other product-batch IPC samples ($s > 1$). Similarly to the second CRISP-DM iteration model, this ML model is trained with all eight inputs (including s, the sample sequence number). The proposed two-stage model ($\widehat{y}_{2\alpha\beta}$) is shown in Fig. 1.

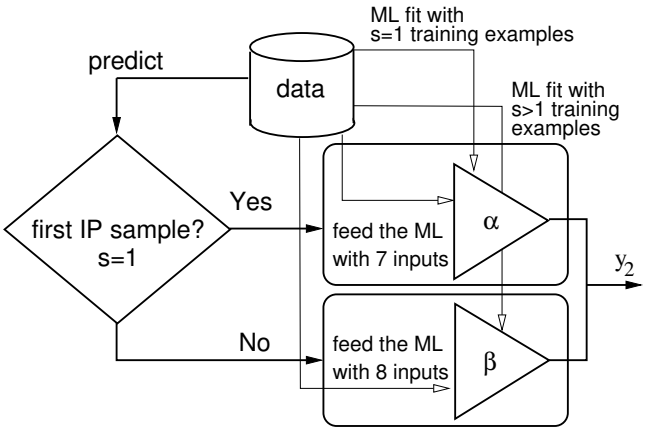

Fig. 1. Schematic of the proposed two-stage ML prediction model ($\widehat{y}_{2\alpha\beta}$).

3.4 Evaluation

The collected data was divided into three main sets, by using a chronological order. The last 20 weeks of data (total of 5,110 examples) was kept out of the initial ML experiments. The goal is apply this additional unseen data in a more realistic evaluation, provided by a Rolling Window (RW) validation [24] that is executed for the best ML regression approach. The remaining and oldest 21,501 examples (not used as test set by the RW) were further divided into training and test sets (holdout split) [20]. The time ordered Holdout Split (HS) was used to compare the three distinct main regression approaches (from the CRISP-DM iterations). The training data included the oldest 15,050 examples (around 70%). As for the HS test set, it included 6,451 instances.

Regarding the RW, it was set using a fixed training window with six months of data and a weekly testing of the ML models, in a total of 20 iterations. In the first iteration, at the first Sunday, the ML was trained with the last six months of historical data. Then, the model was used to perform sample arrival predictions for the incoming week (fixed test size of seven days). In the second iteration, executed at the second Sunday, the training window was updated by discarding one week of the oldest data and adding the previous week examples, allowing to update (retrain) the ML model, which then predicted the next week sample arrival times, and so on.

In this work, we adopt two popular regression error measures: RMSE and Mean Absolute Error (MAE). We also use the Acc@T metric, which is more easily understood by the business analysts, since it measures the percentage of examples accurately predicted when assuming an absolute error tolerance of T. A quality regression model should provide low RMSE and MAE values and also a high accuracy for a small T value. The Acc@T concept allows to compare the predictive performance of different regression modes in a single graph, as proposed in [1] with the Regression Error Characteristic (REC) curves, which plot in the y-axis the Acc@T for different T values (x-axis). The overall quality (for distinct T values) can be measured by computing the Area of REC (AREC) curve when assuming a maximum tolerance of $T_{\max}$ (in %).

4 Results

Table 2 presents the test data errors, in terms of the RMSE error measure, for the HS evaluation and when comparing the two y_2 prediction strategies: $\widehat{y}_2$, executed during the second CRISP-DM iteration; and $\widehat{y}_{2\alpha\beta}$, explored in the third CRISP-DM iteration. The RMSE values confirm that for both prediction strategies, it is more difficult to predict the arrival of the first IPC sample ($s = 1$) than the arrival of the remaining samples ($s > 1$). It is interesting to notice that by specializing a learning model for each of these IPC sample types, as executed in the third CRISP-DM iteration ($\widehat{y}_{2\alpha\beta}$), a substantial error reduction is achieved for both sample types ($s = 1$ and $s > 1$).

Table 2. Test data holdout results for $s = 1$ and $s > 1$ IPC sample arrival (best values in **bold**).

	RMSE	
Approach	$s = 1$	$s > 1$
$\widehat{y}_2$	209.9	188.9
$\widehat{y}_{2\alpha\beta}$	**124.8**	**41.3**

The full comparison of the aggregated HS results, assuming all IPC samples, is shown in Table 3, which contains: the evaluation method used (**Eval.**); the best model selected using the AutoML procedure (**Model**); and several predictive performance measures. The AREC was computed by using a maximum tolerance of $T_{\max=16}$ time units. All performance measures confirm that the best predictive model was achieved by $\widehat{y}_{2\alpha\beta}$, while $\widehat{y}_1$ obtained better results than $\widehat{y}_2$. When compared with $\widehat{y}_1$, $\widehat{y}_{2\alpha\beta}$ achieved a substantial predictive improvement: RMSE – reduction of 46.8 points; MAE – difference of 14.1 points; and AREC – increase of 10 percentage points. As for the ML algorithms, the AutoML selected GBM and SE as the best performing models when using the 10-fold internal cross-validation (applied over training data). The $\widehat{y}_{2\alpha\beta}$ uses GBM for predicting the arrival times of the $s = 1$ samples and SE for the other ones.

Fig. 2 complements the HS results by showing the respective REC curves for the three main regression approaches. The plot confirms that for most of the low tolerance range (x-axis), $\widehat{y}_{2\alpha\beta}$ provides a higher classification accuracy, resulting in an overall higher AREC. Indeed, the proposed two-stage ML model can predict correctly 37%, 59% and 70% of the samples for low tolerance values of $T = 1$, $T = 2$ and $T = 4$, a value that increases to 85% when the tolerance is increased to $T = 16$ time units.

Table 3. Test data results (best HO values in **bold**).

					Acc@T					
Approach	**Eval.**	**Model**	**RMSE**	**MAE**	**AREC**	**T=1**	**T=2**	**T =4**	**T =8**	**T =16**
$\widehat{y}_1$		GBM	98.0	27.0	61%	28%	45%	56%	66%	76%
$\widehat{y}_2$	HO	SE	190.3	112.1	6%	1%	1%	3%	5%	12%
$\widehat{y}_{2\alpha\beta}$		α:GBM;β:SE	**51.2**	**12.9**	**71%**	**37%**	**59%**	**70%**	**77%**	**84%**
$\widehat{y}_{2\alpha\beta}$	RW	α:GBM;β:SE	37.5	11.4	71%	38%	56%	69%	76%	85%

To estimate how the selected model ($\widehat{y}_{2\alpha\beta}$) would behave in a real environment setting, we tested it under a RW evaluation. The results for all 20 week iterations are shown in terms of the last row of Table 3 and show consistency when compared with the HS evaluation. In effect, the same AREC value is achieved (71%), while the RMSE and MAE values are slightly lower (RMSE of 37.5 and MAE of 11.4). This is an interesting result, since the RW evaluation used more recent test data, not seen when comparing the HS results. The obtained results were presented to the business domain experts, which considered them very positive, encouraging the incorporation of the two-stage prediction model into a friendly dashboard that included several business indicators to support the laboratory management decisions. To facilitate the visualization, the dashboard was designed to provide different granularity levels (hourly, daily or monthly) for the sample arrival prediction. For demonstrative purposes, Fig. 3 plots the real and predicted values when assuming a daily aggregation of the IPC sample arrival for a particular chemical laboratory and for the entire RW testing time period. Due to business privacy issues, the scale of the y-axis is omitted from the graph. Fig. 3 shows that the predictions are very close to the real values, denoting a high quality fit of the prediction model.

5 Conclusions

This paper addresses the non-trivial task of predicting the arrival of In-Process Control (IPC) samples at chemical laboratories for quality testing. To solve this task, we implemented the CRoss-Industry Standard Process for Data Mining (CRISP-DM) methodology, under three iterations, each focusing on a different regression approach. During the data understanding and preparation CRISP-DM

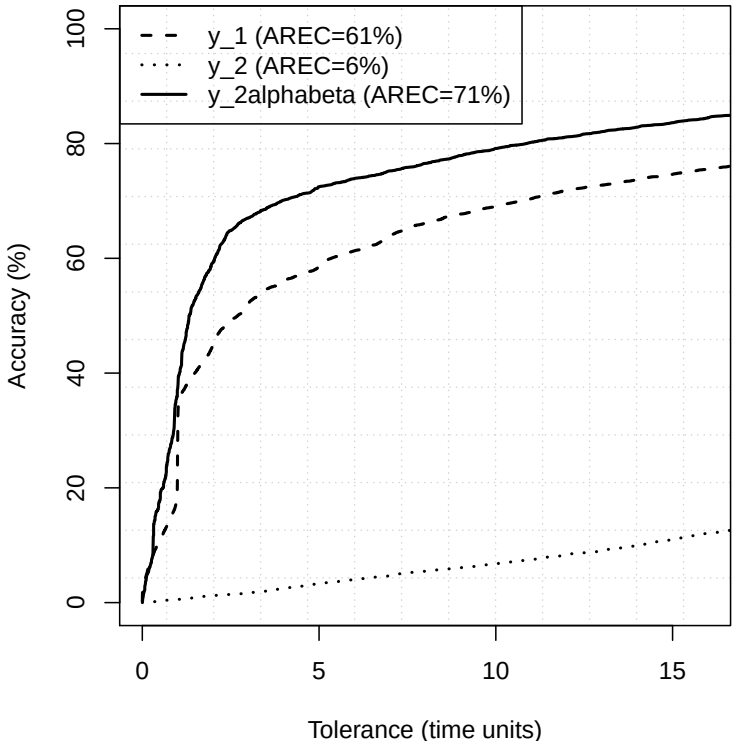

Fig. 2. Holdout REC curves for the three regression approaches.

stages, we collected recent data from a chemical company, resulting in 26,611 sample arrival examples related with a three-year time period. As for the modeling stage of CRISP-DM, we employed an Automated Machine Learning (AutoML) procedure, to automatically select and configure the best model when exploring six state-of-the-art ML algorithms.

Several experiments were held. Using a time ordered Holdout Split (HS), we compared the three main regression approaches: $\widehat{y}_1$ - predict the time lag between the arrival of two consecutive samples (y_1), executed in the first CRISP-DM iteration; $\widehat{y}_2$ - predict the time lag between starting the production of the sample and its arrival to the laboratory (y_2), explored in the second CRISP-DM iteration; and $\widehat{y}_{2\alpha\beta}$ - a two-stage ML model to predict y_2, developed in the third CRISP-DM iteration. For all predictive performance measures, the best results were achieved at the two-stage ML model, which obtained interesting results (e.g., it can accurately predict 70% of the examples under a tolerance of $T = 4$ time units). The selected two-stage ML model ($\widehat{y}_{2\alpha\beta}$) was further evaluated using a realistic Rolling Window (RW) procedure, which considered 20 weeks of unseen data. A similar predictive performance was achieved, when compared with the HS results, showing that the proposed two-stage ML model is robust for the

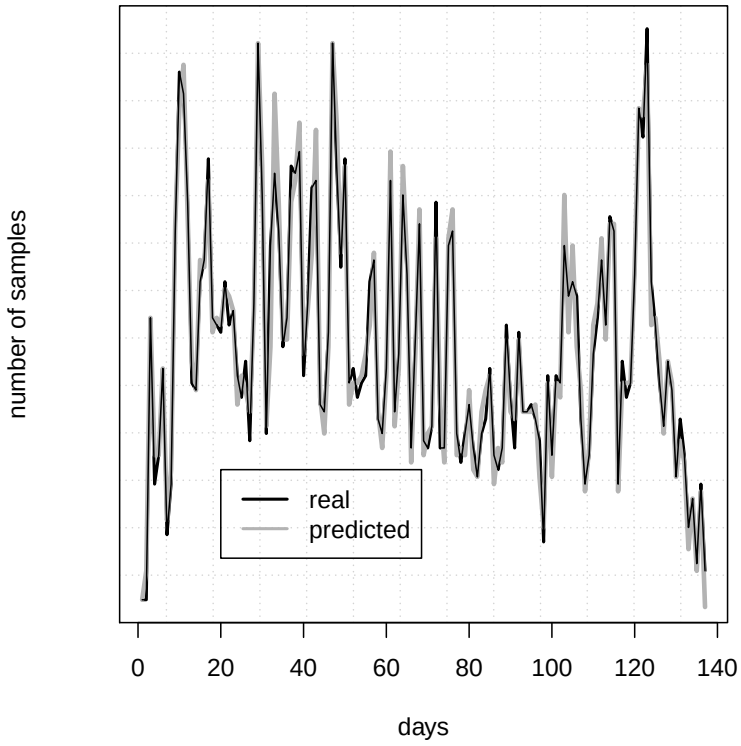

Fig. 3. Daily sample arrival values and $\hat{y}_{2\alpha\beta}$ predictions for the rolling window test data.

analyzed chemical company. In effect, the ML model was incorporated into a friendly dashboard prototype, obtaining a valuable feedback from the chemical laboratory managers.

In future work, we intend to apply the two-stage model to predict the arrival of other types of samples (e.g., raw material). Moreover, we intend to further explore the deployment stage of CRISP-DM, by better integrating the proposed model in a decision support system tool. For instance, by using the predictions to directly optimize the laboratory human resources and instruments.

Acknowledgments

This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020. The authors also wish to thank the chemical company staff involved with this project for providing the data and also the valuable domain feedback.

References

1. Bi, J., Bennett, K.P.: Regression error characteristic curves. In: Fawcett, T., Mishra, N. (eds.) Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA. pp. 43–50. AAAI Press (2003), http://www.aaai.org/Library/ICML/2003/icml03-009.php
2. Breiman, L.: Stacked regressions. Machine Learning **24**(1), 49–64 (1996). https://doi.org/10.1007/BF00117832
3. Caetano, N., Cortez, P., Laureano, R.M.S.: Using data mining for prediction of hospital length of stay: An application of the CRISP-DM methodology. In: Cordeiro, J., Hammoudi, S., Maciaszek, L.A., Camp, O., Filipe, J. (eds.) Enterprise Information Systems - 16th International Conference, ICEIS 2014, Lisbon, Portugal, April 27-30, 2014, Revised Selected Papers. Lecture Notes in Business Information Processing, vol. 227, pp. 149–166. Springer (2014). https://doi.org/10.1007/978-3-319-22348-3_9
4. Canizo, M., Onieva, E., Conde, A., Charramendieta, S., Trujillo, S.: Real-time predictive maintenance for wind turbines using big data frameworks. In: 2017 IEEE International Conference on Prognostics and Health Management, ICPHM 2017, Dallas, TX, USA, June 19-21, 2017. pp. 70–77. IEEE (2017). https://doi.org/10.1109/ICPHM.2017.7998308
5. Charest, M., Finn, R., Dubay, R.: Integration of artificial intelligence in an injection molding process for on-line process parameter adjustment. In: 2018 Annual IEEE International Systems Conference, SysCon 2018, Vancouver, BC, Canada, April 23-26, 2018. pp. 1–6. IEEE (2018). https://doi.org/10.1109/SYSCON.2018.8369500
6. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R. (eds.) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. pp. 785–794. ACM (2016). https://doi.org/10.1145/2939672.2939785
7. Cook, D.: Practical machine learning with H2O: powerful, scalable techniques for deep learning and AI. " O'Reilly Media, Inc." (2016)
8. Cortez, P.: Modern optimization with R. Springer (2014)
9. Darwiche, A.: Human-level intelligence or animal-like abilities? Commun. ACM **61**(10), 56–67 (2018). https://doi.org/10.1145/3271625
10. Deal, J.: The ten most common data mining business mistakes (Jun 2013), https://www.elderresearch.com/most-common-data-science-business-mistakes
11. Ferreira, L., Pilastri, A., Martins, C., Santos, P., Cortez, P.: An Automated and Distributed Machine Learning Framework for Telecommunications Risk Management. In: Proceedings of the 12th International Conference on Agents and Artificial Intelligence, ICAART 2020, Volume 2, Valletta, Malta, February. pp. 99–107. SciTePress (2020)
12. Feurer, M., Klein, A., Eggensperger, K., Springenberg, J.T., Blum, M., Hutter, F.: Efficient and robust automated machine learning. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada. pp. 2962–2970 (2015)
13. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine Learning **63**(1), 3–42 (2006). https://doi.org/10.1007/s10994-006-6226-1
14. Gibert, K., Izquierdo, J., Sànchez-Marrè, M., Hamilton, S.H., Rodríguez-Roda, I., Holmes, G.: Which method to use? an assessment of data mining methods in environmental data science. Environmental modelling & software **110**, 3–27 (2018)

15. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: data mining, inference, and prediction. Springer (2009)
16. Kammergruber, R., Robold, S., Karliç, J., Durner, J.: The future of the laboratory information system–what are the requirements for a powerful system for a laboratory data management? Clinical Chemistry and Laboratory Medicine (CCLM) **52**(11), 225–230 (2014)
17. Liulys, K.: Machine learning application in predictive maintenance. In: 2019 Open Conference of Electrical, Electronic and Information Sciences (eStream). pp. 1–4. IEEE (2019)
18. Moro, S., Laureano, R., Cortez, P.: Using data mining for bank direct marketing: An application of the crisp-dm methodology. In: Proceedings of European Simulation and Modelling Conference-ESM'2011. pp. 117–121. EUROSIS-ETI (2011)
19. Sala, D.A., Jalalvand, A., Deyne, A.V.Y., Mannens, E.: Multivariate time series for data-driven endpoint prediction in the basic oxygen furnace. In: Wani, M.A., Kantardzic, M.M., Mouchaweh, M.S., Gama, J., Lughofer, E. (eds.) 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018, Orlando, FL, USA, December 17-20, 2018. pp. 1419–1426. IEEE (2018). https://doi.org/10.1109/ICMLA.2018.00231
20. Schorfheide, F., Wolpin, K.I.: On the use of holdout samples for model selection. American Economic Review **102**(3), 477–81 (2012)
21. Shrouf, F., Ordieres, J., Miragliotta, G.: Smart factories in industry 4.0: A review of the concept and of energy management approached in production based on the internet of things paradigm. In: 2014 IEEE International Conference on Industrial Engineering and Engineering Management. pp. 697–701 (Dec 2014). https://doi.org/10.1109/IEEM.2014.7058728
22. Skobelev, D., Zaytseva, T., Kozlov, A., Perepelitsa, V., Makarova, A.: Laboratory information management systems in the work of the analytic laboratory. Measurement Techniques **53**(10), 1182–1189 (2011)
23. Spendla, L., Kebisek, M., Tanuska, P., Hrcka, L.: Concept of predictive maintenance of production systems in accordance with industry 4.0. In: 2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI). pp. 000405–000410. IEEE (2017)
24. Tashman, L.J.: Out-of-sample tests of forecasting accuracy: an analysis and review. International Journal of Forecasting **16**(4), 437 – 450 (2000). https://doi.org/10.1016/S0169-2070(00)00065-0
25. Wirth, R., Hipp, J.: Crisp-dm: Towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. pp. 29–39 (2000)