Implementing Federated Learning and Explainability Techniques in Regression Models to Increase Transparency and Reliability*

José RIBEIRO[1]*, Ricardo SANTOS[1], Cesar ANALIDE[2], Fábio SILVA[1]

[1] CIICESI, ESTG, Politécnico do Porto, Portugal
jfr@estg.ipp.pt (**Corresponding author*), rjs@estg.ipp.pt, fas@estg.ipp.pt

[2] ALGORITMI Research Centre / LASI, University of Minho, Campus of Gualtar, 4710-057, Braga, Portugal
analide@di.uminho.pt

Abstract: Interpreting and explaining machine learning models is fundamental for ensuring their transparency and credibility. Explainability techniques such as SHAP and LIME have been developed for this purpose, allowing a deeper understanding of the decisions of AI models. This extended study explores the implementation of a federated learning system for further improving machine learning algorithms. Federated learning allows model training on decentralised data while preserving data privacy and security. The aim of using methods such as SHAP and LIME is to increase the transparency of AI systems and user trust in these systems. These approaches were first implemented in an initial project focused on predictions related to the production time and project management in industry. The obtained results show that federated learning, combined with explainability methods, not only improves the forecast accuracy, but also strengthens the trust among stakeholders. A possible future direction would be to also apply these strategies to other processes in the field of Industry 4.0 and manufacturing, in order to promote the adoption of more comprehensible and innovative intelligent systems.

Keywords: Machine Learning, SHAP, LIME.

1. Introduction

At the intersection of data analysis, machine learning and industry (Ribeiro & Ramos, 2023), a deep understanding of the decisions made by complex models is imperative to increase the reliability and effectiveness of commercial applications. This study builds on previous research in which data was used to explore the decisions of "black box" algorithms for a regression problem. The proposed approach focused on applying advanced interpretability methods to machine learning models. This study utilised techniques such as SHAP (SHapley Additive exPlanations) (Kraus & Feuerriegel, 2019; Gosiewska, Kozak & Biecek, 2021) and LIME (Local Interpretable Model-Agnostic Explanations) (Wang et al., 2023). These techniques were strategically chosen to provide in-depth information on the regression neural network's decisions.

In the context of the initial project, which was centred on competitiveness in industry, the emphasis was on using Artificial Intelligence (AI) techniques to ensure tangible forecasts and to strengthen the trust between the algorithms and the company. This involved creating a forecasting model using ML methods to anticipate the duration of production time for each stage, as well as developing an automatic project management system. The significant progress made in the framework of the initial project facilitated the implementation of explainability techniques in those algorithms. These techniques not only highlight the most influential variables in the predictions (Li et al., 2024), but also provide a detailed insight into the AI decision-making process (Schmitz et al., 2023). Interpretation and explanation have become essential elements for maximising the benefits of this implementation (Wang et al., 2023). The applied techniques have enabled an in-depth understanding of the weight of decisions, contributing to a more informed and transparent decision-making (Burgas et al., 2018), which is fundamental in a scenario where trust, efficiency and competitiveness play central roles (Hjort et al., 2023).

Building on the foundation laid by the previous research such as that of Ribeiro et al. (2024), this extended study explores the implementation of a federated learning system to further improve machine learning algorithms. Federated learning is an innovative approach that allows AI models to be trained on decentralised data while preserving data privacy and security (Kairouz et al., 2021). This technique distributes model training across multiple data sources, allowing each participant to contribute to building more robust models without the need to share sensitive data.

The introduction of federated learning in this context is aimed not only to increase the accuracy

* This article represents an extension of the conference paper: "Exploring Transparency in Decisions of Artificial Neural Networks for Regression", presented at the WorldCIST'24 Conference.

https://doi.org/10.24846/v33i4y202402 ICI Bucharest © Copyright 2012-2024. All rights reserved

and generalisability of predictive models, but also to improve transparency and trust in AI systems. By integrating this approach, regression models are expected to benefit from a greater diversity of data and, consequently, to achieve a better predictive capacity and explainability. This study will investigate the effectiveness of federated learning in optimising the existing algorithms and overcoming the limitations identified in previous research, providing valuable insights with regard to its application in companies where data privacy and prediction accuracy are important.

The remainder of this paper is structured as follows. Section 2 covers the literature review. Section 3 presents a case study, including the study of explainability algorithms, the federated architecture and the results obtained for the SHAP and LIME techniques. Finally, Section 4 sets out the main conclusions.

2. Literature Review

Explainable AI (XAI) algorithms play a fundamental role in providing transparency and interpretability with regard to the decisions of artificial intelligence models. SHapley Additive exPlanations (SHAP) (Kraus & Feuerriegel, 2019) is a common approach for interpreting machine learning models, providing insights into the importance of each feature for both global and local predictions (Schmitz et al., 2023). Based on the cooperative game theory, this technique calculates the Shapley value for each feature, representing its expected contribution to model predictions (Li et al., 2024). Applicable to various model types, SHAP (Gosiewska, Kozak & Biecek, 2021) offers explanations at different levels of detail, from summary graphs to understand overall feature importance to local charts explaining individual predictions. SHAP (Hjort et al., 2023) is a powerful method used in various machine learning tasks, contributing to transparency and confidence by detailing how each feature influences model decisions.

Local Interpretable Model-Agnostic Explanations (LIME) (Wang et al., 2023) is a popular approach for interpreting machine learning models. Similarly to SHAP, LIME is model-agnostic and applicable to different model types. It generates local explanations by adjusting a model's behavior with regard to predictions, making models simpler and more interpretable. The methodology perturbs input data around the instance of interest, observing the resulting changes in the model's output. Despite its flexibility and applicability to any model, including deep learning models, LIME features computational limitations, especially for high- dimensional data. Its approximate descriptions provide valuable insights into understanding a model's behavior in specific cases, making it a powerful tool for interpreting and explaining machine learning models.

Recent machine learning research explores how business profitability can be enhanced through the strategic use of corporate disclosures (Gosiewska, Kozak & Biecek, 2021). The study emphasizes the profitability potential of trading based on corporate disclosures, accounting for transaction costs and liquidity. It underscores the importance of factoring in these considerations when developing investment strategies. The authors utilize SHAP to assess feature importance with regard to corporate disclosures, thereby enhancing the accuracy of machine learning models. SHAP is leveraged to clarify each feature's contribution to model predictions, with graphs illustrating their significance in evaluating machine learning algorithms. The work of Hjort et al. (2023) introduces a modern approach to tree boosting that strikes a balance between predictive accuracy and interpretability, making it ideal for applications relying on fragmented data. The approach is aptly named Locally Interpretable Tree Boosting (LitBoost) and it is applied to predict real estate prices. The interpretability provided by LitBoost values the use of shape functions to create specific local explanations for predictions, illustrating how each attribute contributes to the final prediction.

The study of Schmitz et al. (2023) highlights a new approach, the SafeML framework. Its goal is to enhance the interpretability of machine learning models without compromising their performance. By utilizing complex models to generate innovative features, the framework enables the creation of more transparent and straightforward models. The researchers tested the effectiveness of this framework on various datasets and observed that, in many cases, it achieved a performance which was comparable to that of complex models but with a greater transparency. This suggests that implementing this framework could be the solution for addressing the trade-off challenge between transparency and performance in the field of machine learning.

The study of Wang et al. (2023) presents the development of a machine learning model aimed at identifying patients with inflammatory arthritis and other non-inflammatory conditions during

the screening process. This pioneering study used machine learning to optimize the patient screening process from primary to secondary care. LIME (Local Interpretable Model-Agnostic Explanations) was adopted, providing accurate explanations for each patient-related prediction to enhance the medical decision-making.

The research of Banța et al. (2024) examines the recovery of the automotive industry, highlighting the importance of Industry 4.0 in the process of adaptation and growth after certain global crises, including the COVID-19 pandemic. During this period, production chains suffered significant disruptions, but the adoption of advanced technologies drove their recovery. In 2021 and 2022, the automotive industry not only got back on its feet, but achieved remarkable results, with a record turnover, reflecting the effectiveness of new technological approaches. The transition to Industry 4.0 was central to meeting these challenges, with factories investing in innovation, process automation and technologies such as artificial intelligence, Robotic Process Automation (RPA) and 3D printing. These technological solutions have made it possible to optimize operational flows, improve efficiency and enable product customization, adapting to the demands of an increasingly dynamic market.

The work of Burgas et al. (2018) is an example of a study on optimizing the integrated home care problem through the use of heuristic methods to enhance routing and scheduling decisions. The proposed methods, named ConstructByCaregiver and ConstructbyUser, are carefully designed to find high-quality solutions, considering a wide range of constraints and objectives. Real data from home care providers in Barcelona, Spain, was used for testing these methods, and the results showed that both methods were able to obtain efficient solutions in a reasonable computational time. Additionally, a decision support system was developed to assist in decision-making processes.

The pursuit of explainability, transparency, and interpretability in machine learning models is essential for building trust and understanding. Previously referenced methods play certain roles in this scenario, providing insights that go beyond the black box algorithms (Kraus & Feuerriegel, 2019; Burgas et al., 2018). This approach not only enhances the reliability of predictions but it also strengthens the ability to make informed decisions in various sectors (Wang et al., 2023).

3. Case Study

In the Portuguese business scenario, small and medium-sized enterprises still rely on employees' experience to calculate delivery times without resorting to reliable tools that guarantee accurate estimates. The common practice of processing orders in a First In First Out (FIFO) format, although widely used, may not be the most effective, as it does not consistently deal with possible interruptions, such as a temporary lack of raw materials or a labour shortage.

As such, the aim of this paper was to develop explainability strategies based on a manufacturing process which predicted the time duration for different stages. Recognizing that the transparency of the predictions made may be insufficient to assure trust in the decision process, a new layer of explainability was added to make the product owner understand the basis of time predictions.

Thus, this investigation was grounded in the use of a production database, which included 5 production stages in the manufactoring process, namely: cutting, bending, welding, treatment, and finishing. This database stems from the project presented in (Ribeiro & Ramos, 2023), which was conducted in collaboration with a company in the metalworking sector. This database simulates a manufacturing plant.

Then, a new dataset was introduced that is not only larger, but also more diverse. This new dataset includes all the variables from the original set and also new variables were added that provide a more detailed and accurate view of the production process.

To complement this update, an advanced dataset generator was developed, which is capable of creating synthetic data based on specific statistical parameters such as distribution, and mean and standard deviation. This generator is designed to produce data in a time series format, allowing one to simulate a wide range of production scenarios and to capture the temporal dynamics of manufacturing processes.

By using this dataset generator, it is possible to define detailed metrics for each variable, ensuring that the data generated maintains consistency with the statistical characteristics related to the actual dataset. For example, it is guaranteed that the distribution of the synthetic data follows the

ICI Bucharest © Copyright 2012-2024. All rights reserved

pattern observed for the real data, be it normal, log-normal, exponential, etc. This allows one to accurately simulate variables such as machine operating time and downtime.

It is ensured that the values of the mean and standard deviation of the generated data correspond to the real values, which is important for analysing variables such as the duration of production phases and waiting times. These metrics make it possible to adjust to the variability of data, offering a more realistic representation of production conditions.

In this context, the ability to capture temporal correlations between different variables was implemented, reflecting how past events influence future processes. This is particularly important in a time-series context, where the sequence of data flows and temporal dependence of data directly affects the accuracy of forecasts.

The use of this approach made it possible to create a final dataset that not only expands the amount of available data, but also improves the quality and relevance of the analysed data for this study. This new dataset allows for a deeper and more comprehensive analysis of production processes, providing a solid basis for the development and evaluation of predictive and explainability models in the context of machine learning.

After analysing this dataset, the focus turned to the selected variables, considering their operational relevance and their impact on the proposed model`s performance. The aim of this paper is to deepen the understanding of the behaviour of the artificial neural network model in this specific context, exploring how each algorithm responds and contributes to predictions related to production time in the analysed sector. To improve the explainability of machine learning decisions, the capabilities of the SHAP and LIME algorithms will be employed.

The implementation of advanced explainability methods on this specific data sample should allow the identification of patterns, relationships, and the specific contributions of each variable to the algorithms' decisions. These advanced techniques were applied to the analysed problem, with the aim to gain clear insights into the inner workings of the "black box" and to understand how time each stage is influenced by the input variables.

3.1. The Study of Explainabilty Algorithms

This study focused on the use of SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) explainability techniques applied to a single machine learning algorithm: a neural network. This algorithm was used to solve a regression problem, with the aim of predicting the duration of the various manufacturing stages in a complex industrial process.

The analysed manufacturing steps included cutting, bending, welding, treatment, and finishing. Each of these stages was carefully examined to identify the most influential variables and understand how these variables affect the model predictions. To achieve this, the system was implemented using the Flower library, which supports federative learning. This innovative method allows machine learning models to be trained in a decentralised way, while preserving data privacy by keeping sensitive data on local devices. In this case, each manufacturing step was represented by different local devices, which trained neural network models using the locally available data.

The Artificial Neural Network used is of the MLPRegressor (Multi-layer Perceptron) type, being configured with a single hidden layer made up of 50 neurons, using the ReLU activation function and the Adam optimizer, an adaptive stochastic gradient descent technique. The learning rate is fixed (constant) with an initial value of 0.001, and the L2 regularization is defined by an alpha parameter with the value of 0.0001. The model is evaluated by five-fold cross-validation (K-Fold) and the obtained results are measured in terms of mean squared error (MSE), mean absolute loss (Loss) and accuracy, considering a tolerance of 0.5 to classify predictions as correct. In the context of federated learning, the processes such as those in the metal or metalworking industry, specifically in a metal workshop, can benefit significantly from this type of approach.

With regard to the application of explainability techniques, SHAP was used for quantifying the contribution of each input variable to the model's forecast. SHAP values were generated for each set of variables, making it possible to identify those that influenced the predictions to the highest extent. For example, variables such as "number of simple tasks", "number of complex tasks", "number of employees", "number of machines", "cleaning

difficulty" and "finish type" were analysed to understand their impact on time forecasts.

On the other hand, LIME was used to provide localised interpretations of the model's predictions. This technique generated explanations for specific predictions, helping to understand how input variables influenced individual model decisions. This was particularly useful for identifying and diagnosing situations in which the model could fail.

The graphs generated by the SHAP and LIME techniques provided a detailed and clear view of the decision-making process for the neural network model. These graphs revealed specific areas of each manufacturing stage where optimisations could be made, for promoting the efficiency and improving the quality of the production process.

This study exemplifies the importance of applying explainability techniques to machine learning models, especially in a federated learning system. The use of SHAP and LIME provided valuable insights into the most influential variables at each manufacturing stage and made it possible to transparently explain the model's decision-making process. The combination of advanced technology and data analysis proved relevant in driving significant improvements in complex production environments, ensuring a higher efficiency and quality.

3.2. Federated Learning

Federated learning represents a revolution in the way machine learning models are trained, especially in a context where data privacy is a growing concern. This innovative method allows the decentralised training of models on local devices while preserving data security and privacy (Kairouz et al., 2021). Unlike traditional methods that centralise all data in a single location, federated learning allows the training of machine learning models directly on the devices where the data is generated. Each local participant retains the ownership of his data and trains a model locally. These local models are then combined using aggregation techniques to form a global model, without the raw data being shared between participants (Jas & Dodagoudar, 2023).

A prototype project was developed using Flower, a Python library for federated learning, combining explainability and classification techniques at the edge (on local devices) and centre (at the level of a central server). The Flower library facilitates the implementation of a federated learning system by coordinating decentralised training and the aggregation of local models. In the framework of this prototype, explainability models such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) have been integrated to improve the transparency and understanding of the decisions of the trained models (Smith, Jones & Wang, 2023). This subsection explores how the implementation of federated learning, combined with explainability techniques, can transform data analysis and artificial intelligence in industrial environments.

The process begins with the definition of an initial global model, which is sent to various local devices (to the edge). These devices use their local data to train the model and return only the updated parameters to a central server. Each participating device trains the model using the locally available data. During this process, explainability techniques are applied to identify which characteristics have the greatest impact on model predictions. The parameters of the local models are sent back to the central server, where they are aggregated. This aggregation can be done using the weighted average of the employed parameters or other techniques that take into account the quality of the local models. The global model is updated with the aggregated parameters, incorporating the knowledge acquired from each participating device. This training and aggregation cycle is repeated until the global model reaches a satisfactory performance. After each update cycle, explainability techniques such as SHAP and LIME are applied to the global model to generate insights into the model's decisions. These analyses help identify the influential variables and understand the model's behaviour in different contexts.

Federated learning is a technique that allows collaborative training of machine learning models without the need to centralise the data (McMahan et al., 2017). In this paradigm, the data remains on the local devices, and only the model parameters are shared and aggregated centrally. This method is especially relevant in scenarios where data privacy and security are at stake, such as in Industry 4.0 (Banța et al., 2024) and the Internet of Things (IoT).

In a typical federated learning scenario, the Locksmith process starts with an initial global model, which is sent to several local devices. Each device trains the model using its own data and sends back only the updated parameters to a

ICI Bucharest © Copyright 2012-2024. All rights reserved

central server, as it can be seen in Figure 1. The server then aggregates these parameters to update the global model. This process is repeated until the model converges to a satisfactory solution (Yang et al., 2019).

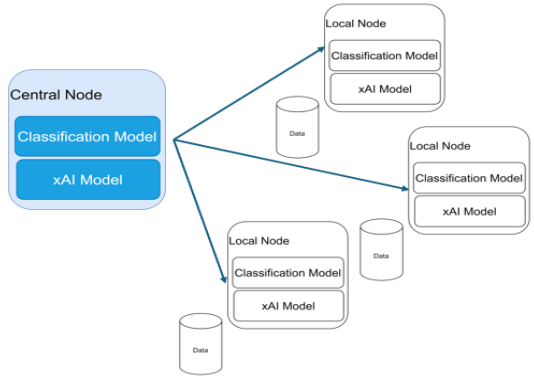

Figure 1. Federated learning scenario

The application of federated learning in industry offers several benefits, such as improving the accuracy of predictive models, protecting data privacy and reducing the need to transfer large volumes of data (Bonawitz et al., 2019). In industrial environments, where safety and operational efficiency are essential, federated learning can provide an effective solution for analysing large volumes of data generated by sensors and IoT devices.

For example, on a production line, different machines can generate data that is used to predict faults or optimise processes. With federated learning, each machine can train a model locally, using its own data, and contribute to a global model that is more robust and accurate. This global model can then be used to predict faults in real time, improving the efficiency and safety of the production line (Khan, Glavin & Nickles, 2023).

The transparency and interpretability of machine learning models are important for their adoption in industrial environments. Explainability techniques such as SHAP and LIME allow users to understand how models make decisions, identifying the most influential variables and providing a clear view of the decision-making process (Ribeiro, Singh & Guestrin, 2016).

When integrated with federated learning, these explainability techniques can significantly improve user confidence and acceptance of AI models. For example, by applying SHAP to a regression model trained in a federated manner, it is possible to identify which variables have the greatest impact on model predictions. This transparency is key to ensuring that the decisions made by the model are understood and accepted by end users (Lundberg & Lee, 2017).

Despite its many benefits, federated learning also features significant challenges, including processed data latency, security issues during the aggregation of local models and the heterogeneity of participating devices (Konečný et al., 2016).

The latency can be a significant problem in environments where processing speed is critical. To mitigate this problem, advanced low-latency data processing techniques can be implemented. These techniques include model compression and the optimisation of communication algorithms between local devices and the central server (Brik et al., 2022).

Data security is a central concern in federated learning. Robust cybersecurity methods, such as homomorphic encryption and secure aggregation, can be used to ensure that data remains protected throughout the training process (Bonawitz et al., 2019). In addition, differential privacy protocols can be implemented to add an extra layer of protection to sensitive data (Dwork, 2006).

The heterogeneity of the participating devices can affect training efficiency. To deal with this diversity, asynchronous learning algorithms can be used, allowing devices with different processing capacities to participate efficiently in training the global model (Li et al., 2020).

Federated learning in combination with explainability techniques has practical applicability in various industrial sectors. Some of these applications are discussed below.

Predictive maintenance in factories and industrial plants is essential to avoid failures and reduce machine downtime. By training predictive models in a federated way, it is possible to use data from different machines and facilities to create a global model that can predict failures with a high accuracy. Explainability techniques, such as SHAP, help to identify which factors contribute to failures, enabling effective preventive action (Nanayakkara, Pokhrel & Li, 2024).

Process optimisation is another area where federated learning can be extremely beneficial. For example, in a factory that produces different types of products, each production line can train a local model to optimise its specific processes. These local models can be aggregated to create a

global model that offers valuable insights into how to optimise the production throughout the factory (Khan, Glavin & Nickles, 2023).

And, as security and privacy in IoT systems represent a central concern, federated learning can be used to train predictive models without compromising data security. Explainability techniques ensure that the decisions made by the employed models are transparent and understandable, promoting user confidence (Hihazi et al., 2024).

The future of federated learning is promising, with a lot of research underway for overcoming the current challenges and expanding its applications. The integration of explainability techniques will continue to play an important role, ensuring that AI models are not only accurate, but also transparent and reliable.

The adoption of federated learning in sectors such as health, finance and transport could revolutionise the way how data is analysed and used, promoting innovation and improving operational efficiency. As technology advances, it is likely that federated learning will be increasingly adopted in industrial environments, which would be driven by the need to protect data privacy and improve the interpretability of AI models (Yang et al., 2019).

Federated learning, combined with explainability techniques, offers a powerful solution for improving the accuracy and transparency of machine learning models in industrial environments. This subsection has explored the fundamentals of federated learning, its benefits, the related challenges and practical applications. By integrating these approaches, various industries can make the most of distributed data, guaranteeing data privacy and security while improving the reliability and effectiveness of AI systems.

3.3. Results

As part of this study, the results obtained by applying an artificial neural network for solving a crucial regression challenge in the industrial context were explored. The central aim was to investigate how specific variables influence the model's predictions during key phases of the production process. Advanced explainability methods were used, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), to uncover the mechanisms underlying the model's decisions.

The choice of a regression neural network is due to its ability to deal with the complex relationships between variables and capture non-linear patterns in industrial data. This approach not only allows for accurate forecasts, but also facilitates the understanding of the important variables that directly impact the efficiency of the production process.

In the first part of the analysis, the SHAP technique was used to visualise and interpret how each variable affects the model's predictions. The normal dot plot revealed valuable insights, with the variable "Number of simple tasks" emerging as the most influential. Significant variations were observed, where negative changes in this variable resulted in a substantial reduction of the forecast accuracy, while positive adjustments contributed to an improvement in this regard. Specific values highlighted negative impacts reaching -75 and positive ones reaching 79 (Figure 2), illustrating the model's important sensitivity to this specific variable.

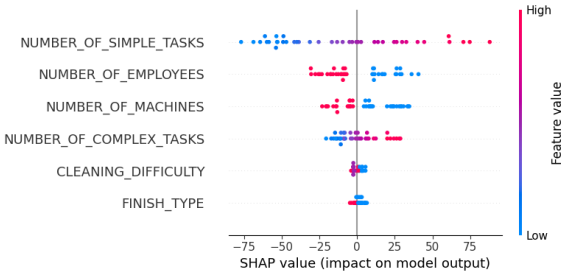

Figure 2. Normal dot plot - SHAP value

Next, the bar chart generated by the SHAP technique is explored, which provides an average view of the influences of the variables on the Artificial Neural Network. However, the explanation of the figures can be expanded to better clarify the obtained results. For example, it was pointed out that the variable "Number of simple tasks" had a considerable impact on the employed model, with a negative influence of -75 and a positive influence of over 75 (Figure 3).

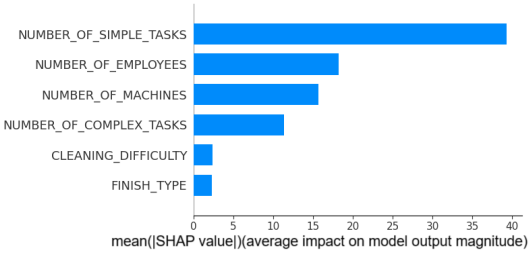

Figure 3. Bar chart – SHAP value

These values indicate that this variable can either significantly decrease or increase the model's output i.e. this variable causes the result of the

ICI Bucharest © Copyright 2012-2024. All rights reserved

algorithm to be close to or far from the real value. The scenario in which the algorithm is applied is related to obtaining information on how long it takes to produce in a certain workstation in the locksmith's workshop, and it is essential to understand how variations in this variable influence the estimated production time.

The results illustrated by the SHAP technique's normal and bar graphs provide an in-depth understanding of the critical variables that influence the neural regression model's predictions. The variable "Number of simple tasks" stands out as a key element for adjustments and optimisations in the production process, which highlights its essential role in improving the efficiency and quality of this process. These insights not only validate the importance of the SHAP approach in interpreting complex models, but also pave the way for future analysis and intervention strategies.

By applying the LIME technique, the influence of the variables related to the employed model was further investigated, focussing on local and detailed interpretations. The graphs revealed that the variable "Number of simple tasks" contributed significantly to the model`s predictions, with a specific negative impact of -70.34. At the same time, the variable "Number of machines" emerged as the most positively influential, with a remarkable value of 8.31 (Figure 4). These findings provide a complementary view of the SHAP results, underlining different aspects of model-variable interactions and enriching the understanding of operational dynamics in the metalworking industry.

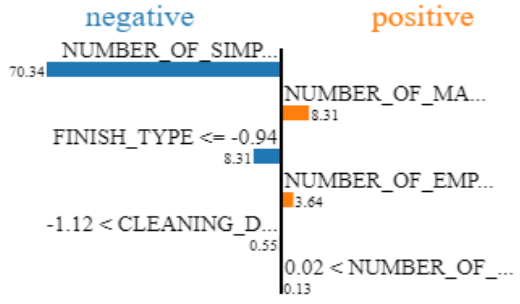

Figure 4. LIME – Graph 1

By applying the LIME technique to explore the influences of the variables specific to the neural regression model, different results that complement the findings for SHAP can be highlighted. The bar chart revealed that the variable "Number of simple tasks" has a significant negative influence on the model's final response, with a specific value of -0.91. This finding suggests that an increase in the number of simple tasks can result in a noticeable reduction in the model's prediction accuracy, emphasising the need to carefully manage this aspect in the context of industrial production. On the other hand, the variable "Number of machines" emerged as the most positively influential, with a value of 1.01 (Figure 5). This indicates that an increase in the number of machines is associated with a substantial improvement with regard to the model's predictions, highlighting it as an important variable for optimising and adjusting operational strategies.

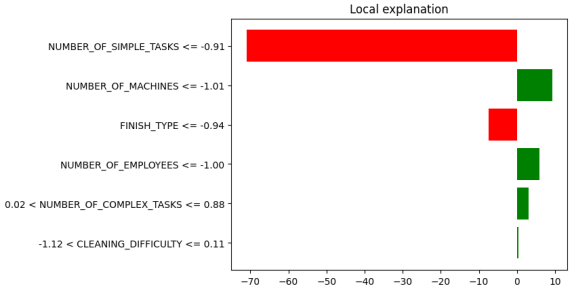

Figure 5. LIME - Graph 2

Integrating the insights related to the SHAP and LIME graphs enabled a more comprehensive view of the influential variables specific to the neural regression model applied to industrial production. Both techniques revealed that the variable "Number of simple tasks" plays a critical role, with distinct negative impacts on the model's predictions. While SHAP highlighted variations between -75 and 79, LIME indicated a negative influence of -0.91, emphasising the consistency of this variable as a central concern for effective production management.

In addition, the variable "Number of machines" consistently emerged as the most positively influential on the obtained results, with significant values of 8.31 for SHAP and 1.01 for LIME. These findings underline the strategic importance of investments and optimisations related to the number of machines to maximise production performance and operational efficiency.

At the same time, the differences between the SHAP and LIME techniques with regard to specific values and graphical representations emphasise the specificity of each approach in interpreting complex models. While SHAP offers a global view of the influences of model variables, LIME provides a local and detailed interpretation, enriching the understanding of dynamic interactions in industrial processes.

This discussion reinforces the need for a multifaceted analysis of the important variables in neural regression models, offering an informed and strategic decisions in industrial management.

In summary, the results obtained using the SHAP and LIME techniques enable a robust and holistic analysis of the influential variables in the neural regression model applied to industrial production. The variable "Number of simple tasks" consistently emerged as a key factor for the model's predictions, highlighting its strategic importance for adjustments and continuous improvements.

4. Conclusion

Model interpretation tools continue to stand out as powerful instruments for understanding and explaining machine learning models. Each method offers a unique perspective on the behaviour of the analysed model, helping to identify the characteristics that determine its predictions. Model interpretation is essential for deciphering the underlying complexities of algorithms, and methods such as SHAP and LIME play important roles in this process.

SHAP ranks the importance of variables, providing a detailed analysis of the contribution of each input variable to the model's predictions. By contrast, LIME is a more localised approach, generating specific interpretations for individual data points. By assigning importance to different characteristics in specific instances, LIME highlights their influences on specific predictions and helps diagnose situations in which the model may fail.

The introduction of federated learning complements and improves these explainability techniques. Federated learning allows the decentralised training of models on local devices, preserving data privacy and increasing data security. This method not only keeps sensitive data secure, but also improves the robustness and generalisability of the trained models in diverse environments. By combining federated learning with explainability techniques, more reliable and transparent models can be obtained, even in contexts where data privacy is fundamental.

Federated learning, when integrated with methods such as SHAP and LIME, can improve the accuracy and reliability of predictions, while providing clear and understandable explanations for stakeholders. The decentralisation of model training allows interpretations and explanations to be adapted to the specific data of each device, providing a more accurate and contextualised view of the model`s predictions.

In the future, these explainability methods could be applied in conjunction with federated learning to different data sets in various contexts, which would broaden the scope of research and enable the validation of results in other contexts, too. The aim is also to compare the results obtained for SHAP and LIME with other explainability techniques, such as Grad-CAM (Gradient-weighted Class Activation Mapping) and additional feature importance techniques, to further the understanding of machine learning models. This integration will provide a robust benchmark for explainability algorithms, promoting trust and transparency for machine learning applications in distributed and privacy-sensitive environments.

Acknowledgements

This work has been supported by national funds through FCT – Fundação para a Ciência e Tecnologia through projects UIDB/04728/2020 and UIDP/04728/2020.

REFERENCES

Banța, V.-C., Sacală, I.-Ș., Țuțui, D., Crețu, R.-F. & Șerban, E. C. (2024) Manufacturing Processes in the Era of Industry 4.0. Case Study: Analysis of a System Architecture in Automotive Industry. *Studies in Informatics and Control*. 33(3), 93-102. doi: 10.24846/v33i3y202409.

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečny, J. K., Mazzocchi, S., McMahan, H. B., Van Overveldt, T., Petrou, D., Ramage, D. & Roselander, J. (2019) *Towards Federated Learning at Scale: System Design*. https://arxiv.org/abs/1902.01046v2 [Accessed 13th December 2024].

Brik, B., Messaadia, M., Sahnoun, M., Bettayeb, B. & Benatia, M. A. (2022) Fog-supported Low-latency Monitoring of System Disruptions in Industry 4.0: A Federated Learning Approach. *ACM Transactions on Cyber-Physical Systems (TCPS)*. 6(2). doi: 10.1145/3477272.

Burgas, L., Melendez, J., Colomer, J., Massana, J. & Pous, C. (2018) N-dimensional extension of unfold-PCA for granular systems monitoring. *Engineering Applications of Artificial Intelligence*. 71, 113–124. doi: 10.1016/J.ENGAPPAI.2018.02.013.

ICI Bucharest © Copyright 2012-2024. All rights reserved

Dwork, C. (2006) Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V. & Wegener, I. (eds.) *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4052. *33rd International Colloquium (ICALP 2006)*, *10-14 July 2006*, *Venice, Italy*. Cham, Switzerland. Springer, pp. 1-12.

Gosiewska, A., Kozak, A. & Biecek, P. (2021) Simpler is better: Lifting interpretability-performance trade-off via automated feature engineering. *Decision Support Systems*. 150, 113556. doi: 10.1016/J.DSS.2021.113556.

Hijazi, N. M., Aloqaily, M., Guizani, M., Ouni, B. & Karray, F. (2024) Secure Federated Learning with Fully Homomorphic Encryption for IoT Communications. *IEEE Internet of Things Journal*. 11(3), 4289–4300. doi: 10.1109/JIOT.2023.3302065.

Hjort, A., Scheel, I., Sommervoll, D.E. & Pensar, J. (2023) Locally interpretable tree boosting: An application to house price prediction. *Decision Support Systems*. 178, 114106. doi: 10.1016/J.DSS.2023.114106.

Jas, K. & Dodagoudar, G. R. (2023) Explainable machine learning model for liquefaction potential assessment of soils using XGBoost-SHAP. *Soil Dynamics and Earthquake Engineering*. 165, 107662. doi: 10.1016/J.SOILDYN.2022.107662.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T. Joshi, G., Khodak, M., Konecný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Florian, T., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H. & Zhao, S. (2021) Advances and Open Problems in Federated Learning. Foundations and Trends® in Machine Learning. 14(1–2), 1–210. doi: 10.1561/2200000083.

Khan, M., Glavin, F. G. & Nickles, M. (2023) Federated Learning as a Privacy Solution - An Overview. *Procedia Computer Science*. 217, 316–325. doi: 10.1016/J.PROCS.2022.12.227.

Konecný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. & Bacon, D. (2016) *Federated Learning: Strategies for Improving Communication Efficiency*. https://arxiv.org/abs/1610.05492. [Accessed 13th December 2024]

Kraus, M. & Feuerriegel, S. (2019) Forecasting remaining useful life: Interpretable deep learning approach via variational Bayesian inferences. *Decision Support Systems.* 125, 113100. doi: 10.1016/J.DSS.2019.113100.

Li, T., Sahu, A. K., Talwalkar, A. & Smith, V. (2020) Federated Learning: Challenges Methods, and Future Directions. *IEEE Signal Processing Magazine*. 37(3), 50-60. doi: 10.1109/MSP.2020.2975749.

Li, Y., Chan, J., Peko, G. & Sundaram, D. (2024) An explanation framework and method for AI-based text emotion analysis and visualisation. *Decision Support Systems.* 178(C), 114121. doi: 10.1016/J.DSS.2023.114121.

Lundberg, S. M. & Lee, S.-I. (2017) A Unified Approach to Interpreting Model Predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), 4-9 December 2017, Long Beach, California, USA*. Red Hook, NY, USA, Curran Associates Inc.. pp. 4768-4777.

McMahan, B., Moore, E., Ramage, D., Hampson, S. & Arcas, B. A. y. (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Singh, A. & Zhu, J. (eds.) *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, *20-22 April 2017, Fort Lauderdale, FL, USA*. pp. 1273–1282.

Nanayakkara, S. I., Pokhrel, S. R. & Li, G. (2024) Understanding global aggregation and optimization of federated learning. *Future Generation Computer Systems*. 159(C), 114–133. doi: 10.1016/J.FUTURE.2024.05.009.

Ribeiro, M. T., Singh, S. & Guestrin, C. (2016) 'Why should i trust you?' Explaining the predictions of any classifier. In: *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *14-18 August 2016, Virtual Event Singapore*. New York, USA, Association for Computing Machinery. pp. 1135–1144.

Ribeiro, J. & Ramos, J. (2023) Forecasting system and project monitoring in industry. In: Omatu, S., Mehmood, R., Sitek, P., Cicerone, S. & Rodríguez, S. (eds.) *Lecture Notes in Networks and Systems (LNNS), vol. 583. Distributed Computing and Artificial Intelligence, 19th International Conference, 13-15 July 2023, L´Aquila, Italy*. Cham, Switzerland. Springer, pp. 249-259.

Ribeiro, J., Santos, R., Analide, C. & Silva, F. (2024) Exploring Transparency in Decisions of Artificial Neural Networks for Regression. In: Rocha, Á., Adeli, H., Dzemyda, G., Moreira, F. & Poniszewska-Marańda, A. (eds) *Lecture Notes in Networks and Systems (LNNS)*, vol. 987. *Good Practices and New Perspectives in Information Systems and Technologies - WorldCIST 2024 (vol. 3*), *26-28 March 2024, Lodz, Poland*. Cham, Switzerland. Springer, pp. 347-356.

Schmitz, H.C., Lutz, B., Wolff, D. & Neumann, D. (2023) When machines trade on corporate disclosures: Using text analytics for investment strategies. *Decision Support Systems.* 165, 113892. doi: 10.1016/J.DSS.2022.113892.

Wang, B., Li, W., Bradlow, A., Bazuaye, E. & Chan, A.T. Y. (2023) Improving triaging from primary care into secondary care using heterogeneous data-driven hybrid machine learning. *Decision Support Systems*. 166, 113899. doi: 10.1016/J.DSS.2022.113899.

Yang, Q., Liu, Y., Chen, T. & Tong, Y. (2019) Federated Machine Learning: Concept and Applications. *ACM Transactions on Intelligent Systems and Technology*. 10(2). doi: 10.1145/3298981.



This is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International License.