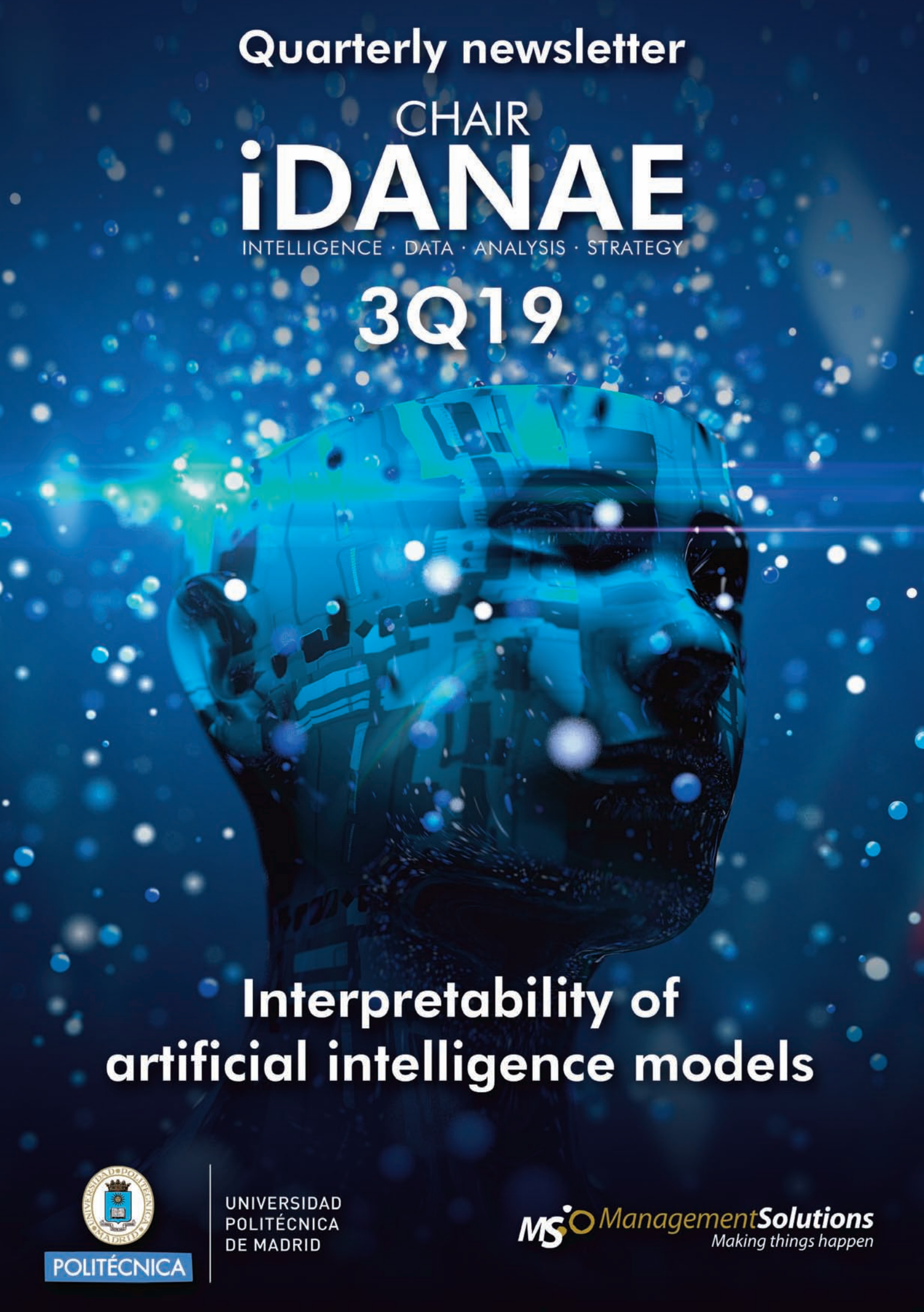
Quarterly newsletter

CHAIR

iDANAE

INTELLIGENCE · DATA · ANALYSIS · STRATEGY

3Q19



Interpretability of artificial intelligence models

POLITÉCNICA

UNIVERSIDAD POLITÉCNICA DE MADRID

ManagementSolutions
Making things happen

Analysis of meta-trends

The iDanae Chair for Big Data and Analytics (where iDanae stands for intelligence, data, analysis and strategy), created within the framework of a collaboration between Universidad Politécnica de Madrid (UPM) and Management Solutions, aims to promote the generation of knowledge, the dissemination and transfer of technology, and the furthering of R&D in the Data Analysis field.

Among the lines of work developed by the iDanae Chair there is the analysis of meta-trends in the field of Analytics. A meta-trend can be defined as a value-generating concept or area of interest in a particular field that will require investment and development from governments, companies and society in the near future.

To identify meta-trends, it is important to analyze public and private investment projects as well as the issues highlighted by organizations, companies and other related stakeholders. The iDanae Chair will conduct active surveillance through observing and monitoring different sources, such as the outcome of different European analytics working groups[1], the strategic plans of the United States Government on artificial intelligence research and development[2], and other relevant international analyses and publications.

For educational and dissemination of information purposes, the findings from this surveillance have been reflected in a list of themes that will be updated as new topics emerge and developed in quarterly reports with the aim of providing insights into specific trends or areas of interest. Some of the topics so far selected for discussion are as follows:

- Interpretability of artificial intelligence models
- Ethical, legal and social implications of artificial intelligence
- Predictability and modeling
- Data augmentation and data democratization
- Data and techniques: a strategic approach to data modeling and strategy
- Practical AI: a development approach.
- Data processing: AI and data protection laws

This first quarterly report is focused on the interpretability of artificial intelligence models.

[1]Big Data Value Association, EU Robotics.
[2]NSTC Committee on AI 2019.

Concept of interpretability

According to BDVA and EU Robotics[3], Artificial Intelligence (AI) is used as an overarching term that covers both digital and physical intelligence, data and robotics, and related smart technologies. AI development includes models and techniques of various levels of complexity, which can lead to a reduction in the human capacity for understanding the underlying models or the obtained results.

The concepts of explainability and interpretability first emerged in response to a need for better understanding of AI models. Whilst both concepts are closely related, and even used interchangeably in the literature, a clear distinction is established by some authors.

For example, in (Gandhi, 2019):

- Interpretability is about the extent to which a cause and effect can be observed within a system, that is, the extent to which you are able to predict what is going to happen, given a change in input or algorithmic parameters.

- Explainability, meanwhile, is the extent to which the internal mechanics of a machine learning system can be explained in human terms. This includes:

  - Explaining the intent behind how the system affects the concerned parties.
  - Explaining the data sources used and how outcomes are obtained.
  - Explaining how inputs in a model lead to outputs.

However, other authors use different definitions of interpretability and explicability. For example:

- Systems are interpretable if their operations can be understood by a human, either through introspection or through a produced explanation (Or Biran, 2017)[4].

- Interpret means to explain or to present in understandable terms. In the context of ML systems, we define interpretability as the ability to explain or to present in understandable terms to a human (Finale Doshi-Velez, 2017).

- We define interpretable machine learning as the use of machine-learning models for the extraction of relevant knowledge about domain relationships contained in data. Here, we view knowledge as being relevant if it provides insight for a particular audience into a chosen domain problem. These insights are often used to guide communication, actions, and discovery (Murdoch, 2019).

[3]BDVA, EU Robotics.
[4]This definition is more similar to the previously definition of explicability.

Therefore, the concept of explicability could be understood in a wider sense, with a more ambitious scope compared to interpretability. The aim of this document is to discuss the concept of interpretability, understood as any built-in technique in an artificial intelligence model that makes it possible to understand the reasons why a specific prediction has been made as well as the relationships between the different variables.

The search for interpretability in machine learning models arises from the growing use of complex models (Hastie, Tibshirani, & Friedman, 2009) such as neural networks, where the complexity of the relationships and calculations as well as the structure and the volume of parameters to be estimated undermine model interpretation. Thus, the aim of interpretation is to understand both the relationships and the results and conclusions.

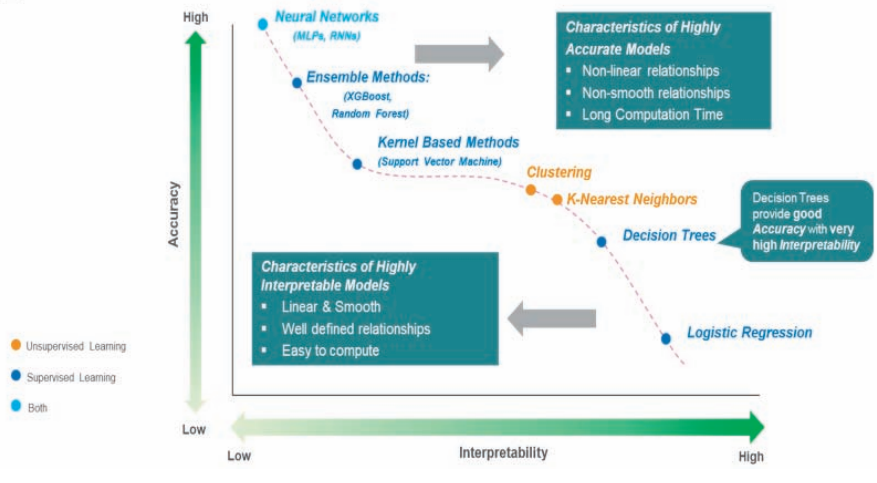
However, in general terms there is an inverse relationship between model interpretability and predictive power: a simpler model is easier to interpret but usually has lower predictive capacity, and vice versa (Exhibit 1).

In general, this trade-off should be analyzed when developing an AI project, considering elements such as the aim of the project, the success metrics or the use that will be made of the result, among others. For example, when interpretability is fundamental in a project, the traditional approach has been to limit the types of machine learning techniques, choosing from between simpler algorithms that can be interpreted from the point of view of their structure or their training; or to simplify complex models so that their operation can be more easily understood. This approach is often advantageous when the relationships in the data are simple (linear). The main techniques used include generalized linear models, decision trees, naive Bayesian classifiers and k-nearest neighbors, which provide high discriminating capacity and/or predictive power.

However, if the use of less interpretable techniques is not wanted, it is possible to search for interpretability from the analysis of the model outputs. The objective in this case is to extract information through several methods, achieving an indirect understanding of the model. For example, small changes in the training model or data can be analyzed and the result observed, or simpler interpretable models that explain the complex model can be constructed (although such models may have worse performance). This document presents some interpretability techniques based on these types of analysis[5].

[5]This techniques are related to the so called 'post-hoc' techniques.

Exhibit 1: Interpretability VS Predictive capacity



Source: Sharayu Rane (2018).

Analysis-based interpretability

Interpretability techniques based on analysis take the trained model as input data (together with the already made predictions), and extract information about the relationships that the model has learned, either training interpretable models using the predictions of the non-interpretable model (known as the construction of a subrogated model), or making perturbations in the input data and studying how the model reacts. These methods are particularly useful when working with complex relationships that need black-box models to achieve reasonable predictive accuracy. In addition, these methods are applicable to different models, and can serve as a measure to compare between the predictions of different models.

This document introduces three of these widely used methods: partial dependence plots (PDP), LIME (Local Interpretable Model-agnostic Explanation) and SHAP (SHapley Additive exPlanations). Other methods not explained in this document include SKATER (Choudhary, Kramer, & team, 2018), ELI5 (Mikhail Korobov, s.f.) and DALEX (Biecek, 2018).

To illustrate these methods, a case study was made using a database of 45,211 entries from a Portuguese bank's marketing campaign to sell a deposit[6]. The campaign was carried out by telephone, with the possibility to call each customer several times in order to sell the product. The aim was to predict whether the customer would subscribe the deposit (variable y). There were variables related to the customer's personal and bank details (age, work, marital status, education, whether customer had any unpaid credit or personal or mortgage loans and the customer's average annual balance), to the current campaign (channel used to communicate with customer, date and duration of the last contact, number of times the customer was contacted) and to the previous campaign (the number of days since the customer was last contacted in the previous campaign, the number of contacts made and the result of the campaign). Exhibits 2-5 shows a visual representation of some of the variables.

Frequency distribution of age.

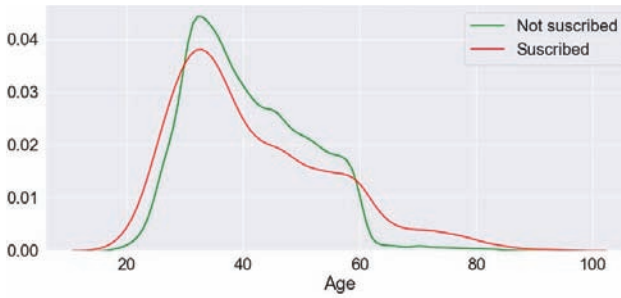


Exhibit 2: Frequency distribution of age for the customers that subscribed the deposit and the customers that did not subscribe it.

Target plot for feature "age".
Average target value through different feature values.

Exhibit 3: Number of contracts and probabilities by age.

Target plot for feature "balance"
Average target value through different feature values.

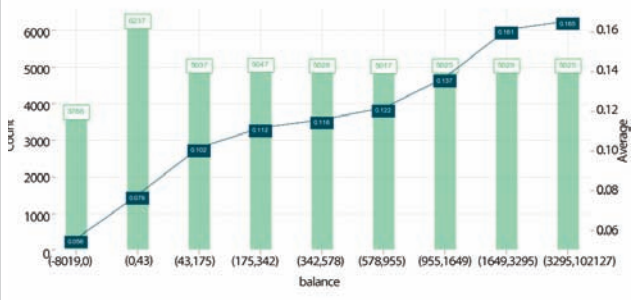


Exhibit 4: Number of contracts and probabilities by balance.

[6]The details on the database can be found in (Moro et alter, 2011)

Target plot for feature "age & balance".
Average target value through different feature value combinations.

Exhibits 5: Contracts by age and balance. The size of the circle shows the absolute number of contracts, whereas the color shows the relative number of contracts.

Partial Dependence Plots (PDP)

Partial dependence plots (Friedman, 2001) show the marginal effect of one or two independent variables on the prediction made, as well as the type of relationship between independent and dependent variables.

This model separates the set of independent variables into two subsets: a subset S with a certain number of variables, and another subset C with the rest of variables. This method works by marginalizing the model's prediction over the distribution of variables in subset C in order to show the dependence between the variables in subset S and the prediction. If S is chosen so that it contains only one or two variables, then it is possible to graphically represent the result in a partial dependence plot.

Individual conditional expectation plots, or ICE plots (Goldstein, 2015), work similarly to PDP plots, but treat each observation individually rather than working with averages. This means that, if there are N observations, this plot will have N lines drawn, instead of a single averaged line as for PDP. This method works better than PDP when there are interactions (two or more variables interacting to generate a new effect) between the independent variables.

Case study: Exhibit 6 shows an ICE and PDP chart for the Balance variable, whereas Exhibit 7 shows the same chart for the Duration variable (duration of the last call). The blue lines are the ICE plots for a subsample of 300 predictions, while the yellow line shows the PDP, which is the average of all ICE plots. A larger balance means the customer is more likely to purchase the deposit product, while a longer call duration also indicates a greater probability that the customer will subscribe the deposit.

PDP for feature "balance".
Number of unique grid points: 10.

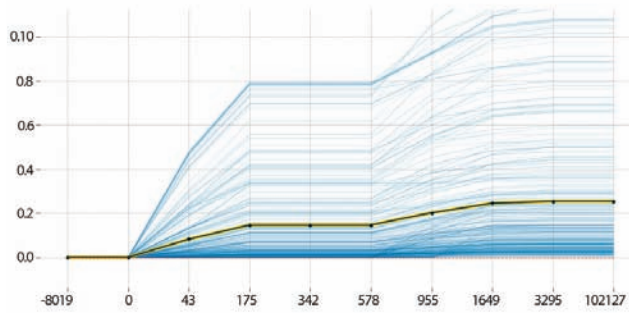


Exhibit 6: PDP (yellow line) and ICE (blue lines) for Balance variable.

PDP for feature "duration".
Number of unique grid points: 10.

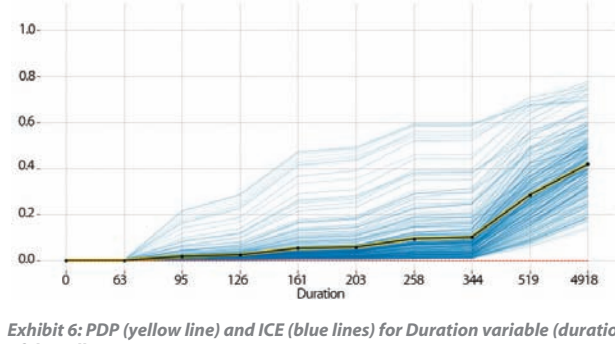


Exhibit 6: PDP (yellow line) and ICE (blue lines) for Duration variable (duration of the call).

LIME

LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro, Singh, & Guestrin, 2016) is a local method that checks the predictions of a model when the input data varies. This is not a purely transparent model, as it provides the explanation after a decision has been made. LIME generates new data composed of modified data and the predictions generated by the model. These new data are then used to train an interpretable model such as those mentioned above: linear models, decision trees, etc. This interpretable model should be a good approximation of predictions locally.

The explanation ξ is defined as a model $g \in G$ where G is a class of potentially interpretable models. Since not all models are simple enough to be easily interpreted in all situations, a measure of complexity $\Omega(g)$ is defined (for decision trees, for example, it is the depth of the tree; for linear regressions, the number of non-zero weights). A proximity measure $\pi_x(z)$ to determine a distance between observations x and z must be defined in order to establish a neighbourhood of x. This allows $L(f,g,\pi_x)$ to be defined as a measure of fidelity in the approximation of f by g in a neighbourhood of x defined by π_x. Therefore, to obtain an interpretation of the prediction of x and having a good local approximation of the model, $L(f,g,\pi_x)$ must be minimized while keeping $\Omega(g)$ low enough to be interpretable by humans.

Case study: Exhibit 7 shows the LIME explanation for one of the predictions. In this specific case, the model predicts that the customer is not going to buy the deposit product, and this prediction is mainly due to the values of the duration variable (duration of the last call), pdays variable (number of days since customer was last contacted in the previous campaign, with -1 being the assigned value if customer has not been contacted), previous variable (contacts made in the previous campaign) and contact variable (channel through which customer is contacted: unknown, phone or mobile phone).

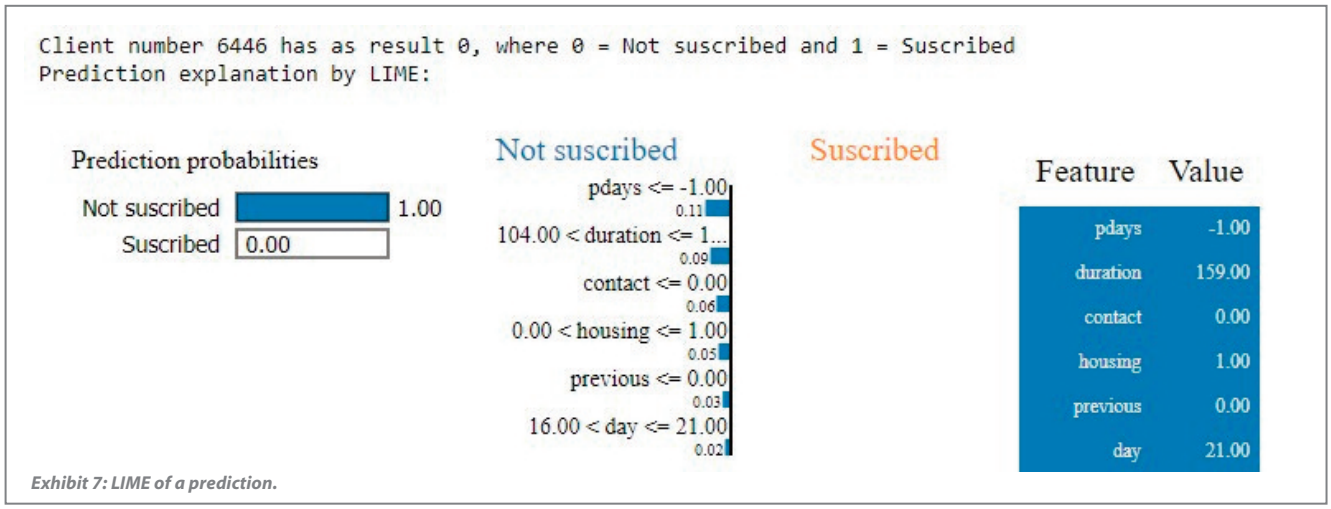


Exhibit 7: LIME of a prediction.

SHAP

SHAP (Scott Lundberg, 2017) is a local method that uses cooperative game theory to interpret the outputs predicted by a model. This method computes the Shapley values (Shapley, 1953), interpreting the independent variables as players that collaborate to receive a payoff, which in this case is the specific prediction made by the model minus the average value of all predictions. The players "share" the payoff based on their contribution, calculated using the Shapley values.

This method requires retraining the model for every possible $S \subseteq F$ subset, where F is the set of independent variables, assigning each variable an importance value that represents its effect on the prediction. For this, a model $f_{(S \cup \{i\})}$ is trained with the study variable present, and another model f_S is trained with the variable withheld. The predictions from both models are compared for the specific input x_S to be predicted.

Since the effect of withholding a variable depends on other variables in the model, these differences are calculated for all possible subsets $S \subseteq F \setminus \{i\}$, the Shapley values being calculated as the weighted average of all possible differences.

Case study: Exhibit 8 shows the same prediction previously explained by LIME. The variables Contact, Housing and Pdays are the biggest contributors to increased probability. It should be noted that SHAP does not show the probability of default but the average contribution of a variable to the prediction (how much the prediction is modified on average considering all subsets with the variable present and with the variable withheld). It is not a probability or a value that indicates the difference in the prediction if that variable is eliminated.

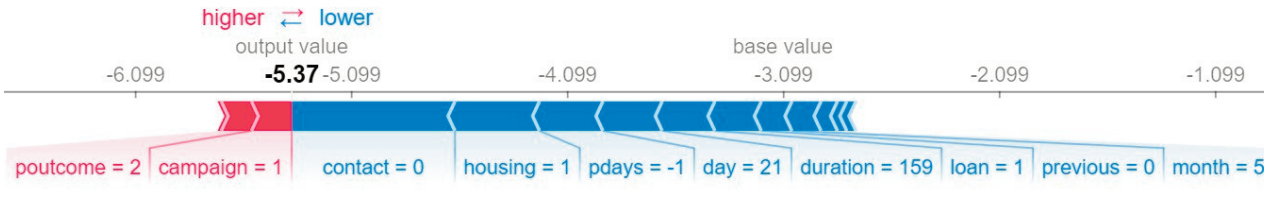


Exhibit 8: Prediction explained by SHAP.

Comparison between PDP, LIME and SHAP

The following table shows a comparison between the three methods.

	PDP	LIME	SHAP
CONCEPT	Marginalizes all variables except one to see how variations in the variable affect the prediction.	Approximate small variations of the prediction, locally, to interpretable models.	Treats the prediction as a cooperative game between the variables in which a payoff (the prediction made) is distributed weighted to its contribution.
ADVANTAGES	The interpretation is very intuitive and causal, besides being simple to implement.	It is not necessary to use all the variables and it works for texts and images. It has predictive capacity for variable environments.	The mathematical foundations behind SHAP make it a solid interpretation theory.
DISADVANTAGES	It only allows one or two variables to be explained at a time. Independence between the variables is assumed.	It is complicated to define the environment of the variables. Explanations can be considerably different across small variations in the variables.	It involves a high computational cost and the result can be misunderstood. All variables are always used and have no predictive ability.

Table 1: Comparison between PDP, LIME and SHAP.

References

BDVA, EU Robotics: "Strategic Research, Innovation and Deployment Agenda for an AI PPP". Consultation Release. June 2019.

Biecek, P. (2018). DALEX: explainers for complex predictive models. The Journal of Machine Learning Research.

Choudhary, P., Kramer, A., & team, d. (2018). Skater: Model Interpretation Library. https://zenodo.org/record/1198885.

Finale Doshi-Velez, K. B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. Retrieved from Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of statistics, 1189-1232.

Gandhi, P. (2019). Explainable Artificial Intelligence. Retrieved from KD Nuggets: https://www.kdnuggets.com/2019/01/explainable-ai.html

Goldstein, A. e. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics 24.1, 44-65.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning. Springer.

M. Lichman, U. o. (s.f.). UCI Machine Learning Repository. Retrieved from UCI Machine Learning Repository: http://archive.ics.uci.edu/ml

Mikhail Korobov, K. L. (s.f.). ELI5. Retrieved from Github: https://github.com/TeamHG-Memex/eli5

Murdoch, W. J. (2019). Interpretable machine learning: definitions, methods, and applications. arXiv preprint arXiv:1901.04592.

Moro et alter, 2011: S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.

NSTC Committee on AI 2019: Committee on AI of the National Science and Technology Council: The national artificial intelligence research and development strategic plan: 2019 update. Executive Office of the President of the United States. June 2019.

Or Biran, C. C. (2017). Explanation and Justification in Machine Learning: A Survey. IJCAI 2017 Workshop on Explainable Artificial Intelligence.

Ribeiro, M. T., Singh, S., & Guestrin, a. C. (2016). Why should I trust you?: Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM.

Scott Lundberg, S.-I. L. (2017). A Unified Approach to Interpreting Model Predictions. NIPS.

Shapley, L. S. (1953). A value for n-person games. Contributions to the Theory of Games.

Sharayu Rane (2018): "The Balance: Accuracy vs Interpretability". Towards Data Science. 2018.

Authors

Ernestina Menasalvas (UPM)
Alejandro Rodríguez (UPM)
Manuel Ángel Guzmán (Management Solutions)
Segismundo Jiménez (Management Solutions)
Carlos Alonso (Management Solutions)

UNIVERSIDAD POLITÉCNICA DE MADRID

The Universidad Politécnica de Madrid is a public-law organization of a multisectoral and multidisciplinary nature that is engaged in teaching, research, as well as science and technology development activities.

www.upm.es

Management Solutions is an international consulting firm whose core mission is to deliver business, risk, financial, organizational and process-related advisory services, with operations in more than 40 countries and a multidisciplinary team of 2,500 professionals working for over 900 clients worldwide.

www.managementsolutions.com

For more information, visit

blogs.upm.es/catedra-idanae/