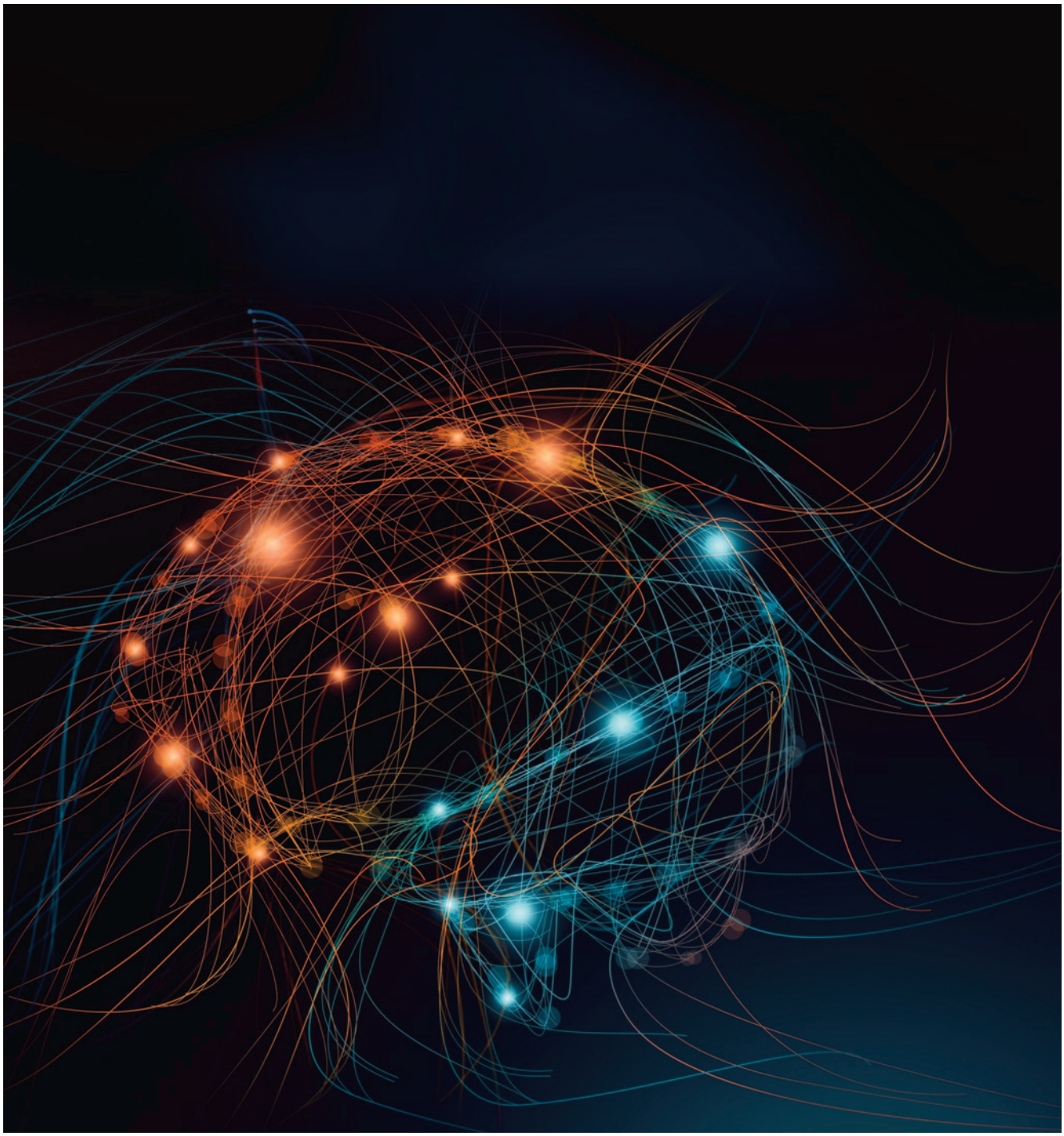
Interpretability use case

"Fools ignore complexity. Pragmatists suffer from it. Some can avoid it. Geniuses remove it"
Alan Perlis[72]



Approach

This section presents a use case for AI interpretability to illustrate how the XAI techniques described in the previous section are applied.

The selected use case addresses the problem of employee retention in an organization, focusing on understanding and explaining the causes that lead employees to leave their jobs. Identifying these factors can enable organizations to take preventive measures and develop strategies to improve job satisfaction and talent retention.

This use case is based on a fictitious dataset generated by IBM and published in Kaggle[73]. This dataset contains information about an organization's employees, including demographic characteristics, data about their job title, and whether they have left the company.

In the year under review, the company has an employee attrition rate of 16%, 6% above the historical average, and is concerned about finding out the causes to be able to develop a remediation plan.

The main variables present in the data set include:

- Level of education (from "high school" to "Ph.D.").
- Satisfaction with the work environment (from "low" to "very high").
- Job involvement (from "low" to "very high").
- Job satisfaction (from "low" to "very high").
- Performance rating (from "low" to "outstanding").
- Satisfaction with labor relations (from "low" to "very high").
- Work/life balance (from "bad" to "optimal").
- Years since last job promotion (numerical variable).
- Monthly salary (numerical variable).
- Years in current job (numerical variable).
- Distance from home to work (numerical variable).
- Number of companies in which the employee has worked (numerical variable).
- Role in current job (categorical variable, includes "Manager", "Director", "Research Scientist", among others).

The focus of this use case was to train and validate different artificial intelligence models to predict employee attrition, using XAI techniques to analyze and understand the behavior and decisions of the selected models.

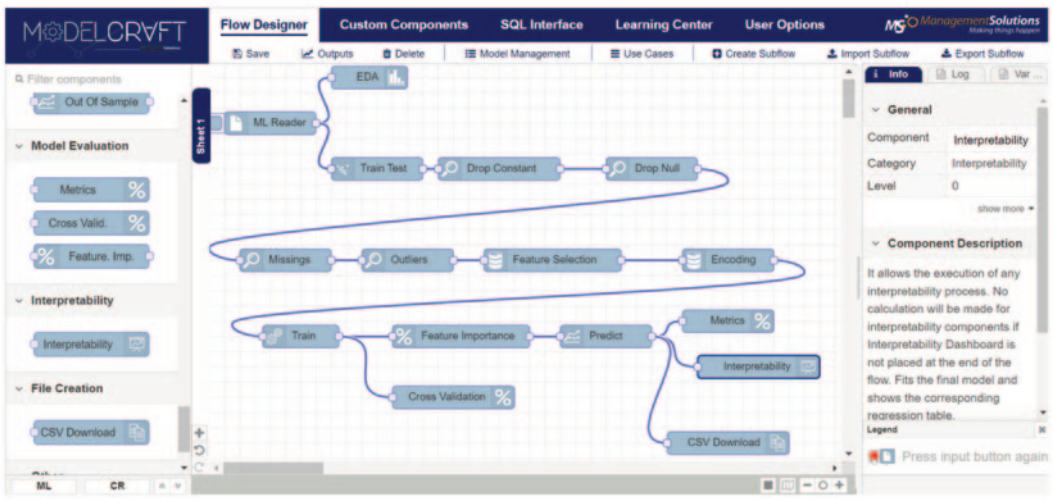
To simplify and streamline the process, the ModelCraft™ component modeling system, which contains multiple relevant AI and XAI techniques, was used. This system allowed the study to be carried out efficiently and without the need to write code.

Throughout the use case, the SHAP, LIME and PDP interpretability techniques were applied to analyze the selected models and understand which variables influence employees' decisions to leave their jobs. In addition, this use case explored how these variables interacted with each other and how they affected different segments of the employee population.

At the end of the use case, the effectiveness and limitations of the interpretability techniques used will be evaluated. There will be a discussion on how the combination of artificial intelligence models and interpretability modules can improve the predictive capability and understanding of the models, thus facilitating data-driven decision making in the business domain.

[72] Alan Jay Perlis (1922-1990), american computer scientist, PhD in Computer Science from MIT and professor at Purdue University, Carnegie Mellon University and the University of California at Berkeley, known for his pioneering work in programming languages and for being the first winner of the Turing Award.
[73] Kaggle (2017). IBM HR Analytics Employee Attrition & Performance.

Figure 13. Modeling flow in ModelCraft™.



Modeling process

The modeling process is carried out in three phases: data engineering, modeling and model interpretability analysis.

1. Data engineering

Data engineering is the initial phase in which the data set is prepared and processed for use in the creation of artificial intelligence models. In this case, the following actions were performed:

- Definition of the scope of application: in this case, the population was taken as all employees who had been on sick leave in the previous two years.
- Data cleansing: data quality was verified and records with missing or inconsistent information were eliminated or corrected.
- Variable transformation: categorical variables were converted into numerical variables using techniques such as one-hot encoding or ordinal encoding. In addition, numerical variables were normalized or standardized when necessary.
- Variable selection: the most relevant variables for predicting employee attrition were identified using variable selection techniques such as Pearson correlation, feature importance in tree-based models or recursive feature elimination.
- Feature engineering: new variables were generated from existing ones to analyze whether they were better predictors of employee turnover, such as "total satisfaction", which was constructed as the sum of the scores of the variables "Satisfaction with the work environment", "Job satisfaction", "Performance rating", "Work-life balance", "Job involvement " and "Satisfaction with labor relations".
- Train-test split: the dataset was divided into two subsets: training and testing. The training subset was used to tune and optimize the artificial intelligence models, while the test subset was used to evaluate the performance and predictive power of the models.

2. Model development

In this phase, different artificial intelligence models were trained and validated using the training subset. Specifically, several of the most common machine learning algorithms and traditional models, such as logistic regression, decision trees, support vector machines, neural networks and random forest, were fitted and compared to select the model with the best performance.

To avoid overfitting and to optimize the hyperparameters of the models, cross-validation and grid or random search techniques were used. In addition, model complexity was given particular consideration when selecting a specific algorithm during training in order to facilitate model interpretation.

For this purpose, a model development flow was generated in ModelCraft™ (Fig. 13).

To select the model with the best predictive power, its performance on the test subset was evaluated using metrics such as accuracy, sensitivity, specificity and area under the ROC curve (AUC-ROC). These metrics allowed us to evaluate the effectiveness of the selected model in terms of its ability to correctly predict employee attrition on previously unseen data.

All things considered, the random forest yields superior performance results, although it poses an interpretability challenge in understanding its predictions. This model has considered 300 decision trees and has yielded an accuracy of 75% and a sensitivity of 84%. Therefore, these are very reliable

predictions and false negatives are rarely obtained. This is relevant for this use case, where the company would foreseeably want to reduce this type of error as much as possible.

3. Interpretability analysis

In this last phase, interpretability techniques were applied to analyze and understand the behavior and decisions of the selected model. Specifically, the objectives of the analysis were:

- To understand which variables were most important in decision making for the organization at a global level, for which purpose a comparison of the importance of each variable was used.
- To understand how changes in the most important variables impact different population ranges.
- To understand model results in specific cases where a certain probability of abandonment is observed.

In this use case, SHAP, LIME and PDP techniques were used to explain how the model made decisions and how the inputs influenced the predictions.

SHAP allowed to obtain local and global interpretability results, which provided an interpretation of the importance of each variable, and LIME allowed us to perform an intuitive analysis of local interpretability that made it possible for us to explain the outcome of the model for each employee based on simpler linear models. As a complement, PDP graphs allowed visualization of how changes in each variable impacted the model's prediction.

Thus, the following distribution of the importance of each variable was obtained (Fig. 14).

In this case, it was observed that the variable with the greatest importance in predicting employee attrition (15.65%) was "overall satisfaction", a synthetic indicator defined as a weighted average of six elements (work environment, suitability of functions and areas to the position, internal rating, work/life balance, relationship with colleagues and supervisors, and employee position and responsibility).

This result was intuitive and showed that the "overall satisfaction" variable was well designed. However, the next three variables by importance (length of service, salary, and distance from home to work) appeared to have a high influence on employee turnover, which collectively doubled that of the "overall satisfaction" indicator.

To understand how each variable was influenced individually, the PDPs were studied (Fig. 15).

In terms of length of service, the trend was reversed after three years: employees with intermediate length of service were, on average, the least likely to leave the company. For overall satisfaction, an intuitive trend was observed: higher satisfaction reported in internal surveys resulted in a lower quit rate.

To complement the previous analysis, LIME was used for a case-by-case analysis of the values of variables influencing the likelihood of certain employees leaving the company. Fig. 16 shows two employees with different quit probabilities obtained using the model. LIME shows an explainability metric representing how good a linear fit it has obtained using the local surrogate model to explain these predictions.

It is interesting to see how the most significant causes of employee abandonment in these two cases do not necessarily correspond to the most influential variables at the global level. While overall satisfaction appears to contribute to explaining the likelihood of employee abandonment in case 1, it does not seem to have a significant impact in case 2, where the

Figure 14. Global interpretability of the random forest model using SHAP, where the Shapley values are used to obtain the importance of the variables.

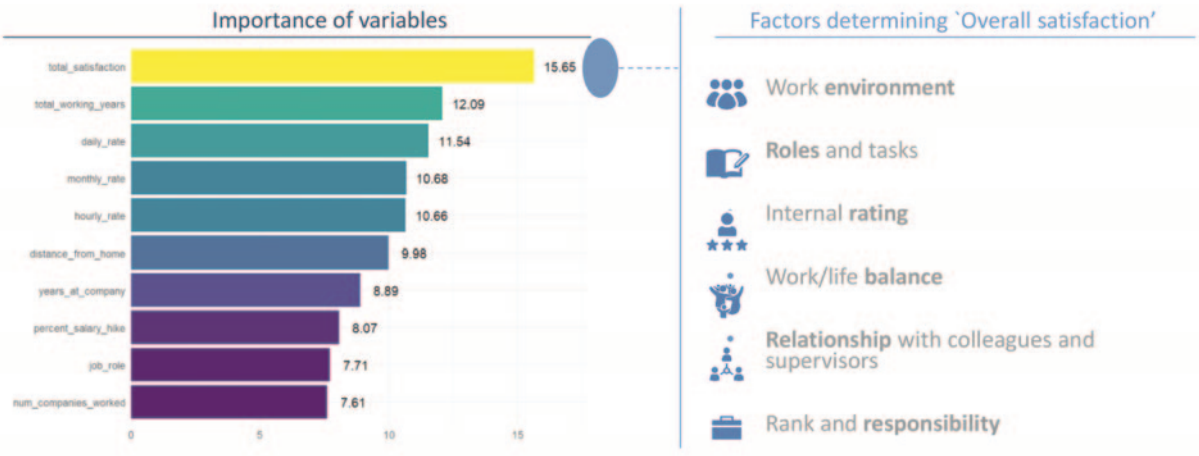
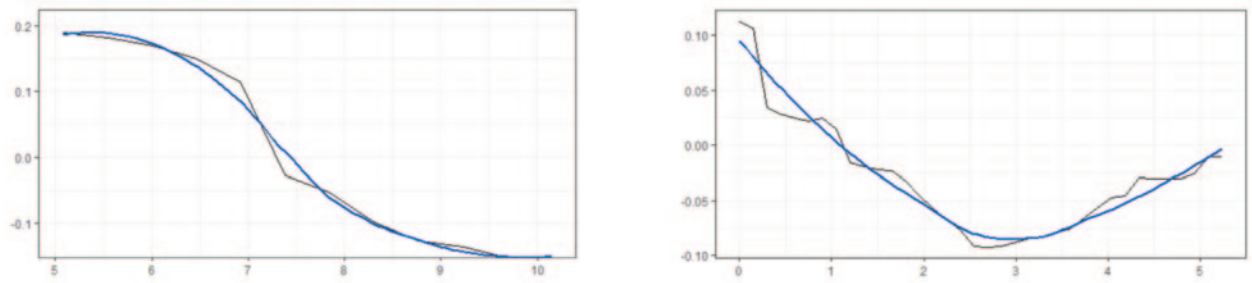


Figure 15. PDP plots for the variables "total satisfaction" and "length of service".



probability of leaving is higher.

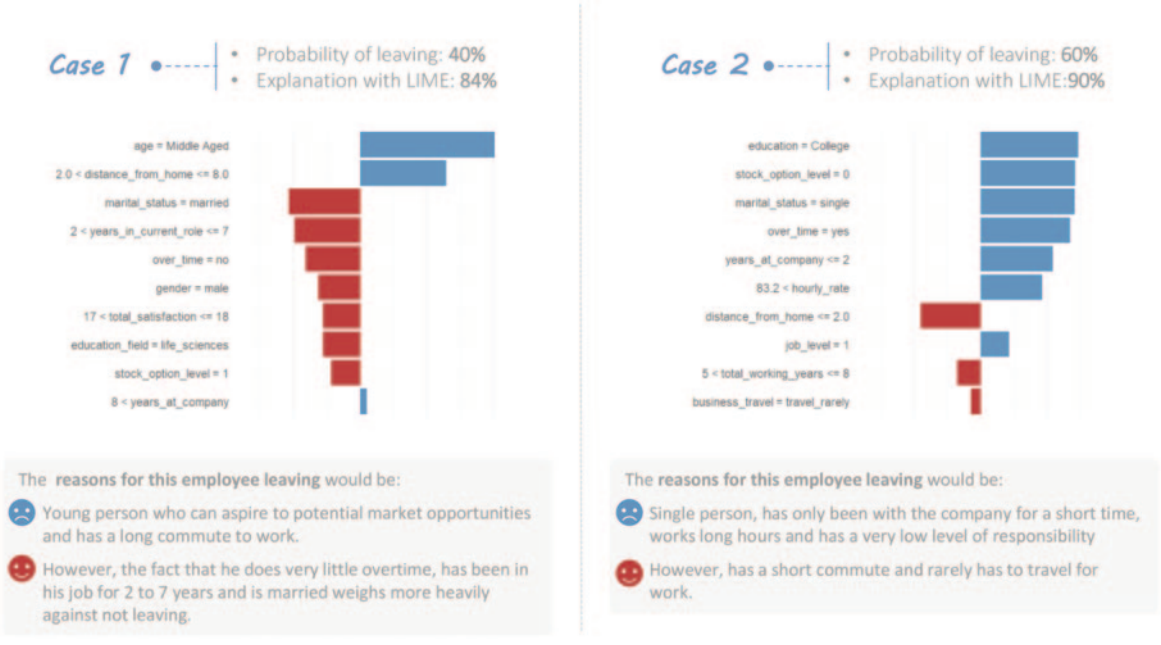
This reflects the difficulties in interpreting this model, which can be generalized to similar models: although overall satisfaction can explain the average probability of employee abandonment, this conclusion is a generalization, as there are individual and group cases in which employee abandonment is explained to a greater extent by other variables.

Conclusions of the use case

Several conclusions and lessons learned can be drawn from this artificial intelligence interpretability use case that may be useful in future uses of AI and XAI models:

- **Model use:** the correct use and interpretation of the model in this case can make it possible to anticipate and prevent employee turnover. Among the uses that can be made of the model is the ability to create different profiles with a propensity to leave and identify the characteristics of these employees in advance to take appropriate measures, which in the long term can contribute to reducing the level of turnover in the organization.
- **Model selection:** the modeling process demonstrated the importance of comparing and validating different machine learning algorithms to select the model with the best predictive capability. In this case, the random forest model proved to be the most suitable for predicting employee attrition.

Figure 16. Local interpretability of the random forest model using LIME.



- **Importance of interpretability:** the use of interpretability techniques, such as SHAP, LIME and PDP, provided a deeper understanding of how the model makes decisions and how inputs influence predictions. This information is crucial to validate the applicability of the model in a real-world context and to ensure that predictions are based on relevant and meaningful features.

- **Influential variables:** the interpretability analysis allowed us to identify the most relevant variables for predicting employee attrition. These variables can be useful in developing retention strategies and improving job satisfaction. In addition, understanding how these variables interact with each other and how they affect different segments of the employee population can enrich the analysis and facilitate data-driven decision making.

- **Practical implementation:** the use case demonstrates the feasibility and usefulness of applying AI and XAI techniques in a realistic scenario, using fictitious data but representative of a business situation. This approach can be adapted to other business contexts and problems, taking advantage of artificial intelligence and interpretability to improve decision making and obtain more efficient and effective results.

- **Constraints:** at the same time, this use case highlighted the constraints and difficulties in the use of post-hoc interpretability techniques. It is important to recognize that interpretability methods are not infallible and may sometimes provide approximate or partial results. Therefore, it is essential to take a critical and rigorous approach when interpreting and validating the outcome of interpretability techniques.

- **Combining AI models and interpretability modules:** this use case shows how the integration of AI models and interpretability modules can improve the predictive capability and understanding of models. This facilitates the adoption of AI-based solutions in business decision making.

- **Continuity in interpretability analysis:** finally, it should be emphasized that interpretability analysis should not be an isolated exercise applied during model development, but should be performed in a continuous, reproducible and reliable manner throughout the life of the model.

In conclusion, this artificial intelligence interpretability use case provided valuable experience in the implementation of AI and XAI techniques in a business context, and shows the potential of AI and interpretability to improve decision making, while revealing the limitations and difficulties associated with these techniques and the need for a critical and rigorous approach when interpreting and validating AI results.