UCLouvain

FACULTE DES SCIENCES

Faculté des sciences

XAI : A Comparative Study of Post-hoc and Intrinsic Explainability Methods for Categorical Tabular Data

Author: **Siméon Godfrin**
Supervisors: **Olivier Caelen, Corentin Vande Kerckhove, Robin Van Oirbeek**
Reader: **Marco Saerens**

Academic year 2025–2026
Master [120] en science des données, orientation statistique

Remerciements Je remercie la LSBA et tous ces professeurs pour la formation reçue.
Je remercie mes promoteurs : Monsieur Rovin Van Oirbeek pour le soutient et la planification de cette tâche; Monsieur Olivier Caelen pour la discipline quotidienne et Monsieur Corentin Vande Kerckhove pour l'accès au donnée.
Je remercie Monsieur Marco Saerens pour la découverte de cette branche du Machine learning qu'est la XAI.

Je remercie ma copine pour le soutient et Guillaume Haquin pour la relecture.

Ceci est la version corrigée de mon mémoire

Contents

1 Introduction **5**

1.1 Contextualization . . . 5

1.2 Organization . . . 6

1.2.1 Defintion of XAI . . . 6

1.2.2 The three axis . . . 8

2 Exploratory Data Analysis, Preprocessing and Models **9**

2.1 Description of the dataset . . . 9

2.1.1 Summary of all variables . . . 14

2.2 Preprocessing . . . 16

2.2.1 Missing data . . . 16

2.2.2 Imputation model . . . 16

2.2.3 Geographical information . . . 17

2.2.4 Categorical variables . . . 17

2.2.5 Highly imbalanced datasets . . . 17

2.2.6 Response variable and people_predict variable . . . 18

2.2.7 Parallel dataset construction for method compatibility . . . 18

2.2.8 Observation alignment via absolute indexing . . . 18

2.2.9 Resolution of inconsistent categorical encodings . . . 18

2.2.10 Final dataset size . . . 19

2.3 The neural network model . . . 20

2.3.1 Model role and setup . . . 20

2.3.2 Architecture . . . 20

2.3.3 Hyperparameter optimization . . . 20

2.3.4 Performance . . . 21

3 Post-hoc **22**

3.1 Introduction . . . 22

3.2 LIME . . . 22

3.3 SHAP . . . 23

3.3.1 Introduction . . . 23

3.3.2 The properties . . . 24

3.4 Diverse Counterfactual Explanations (DiCE) . . . 26

3.4.1 Optimization . . . 27

3.4.2 Choice of the loss . . . 27

3.4.3 Proximity and diversity term . . . 27

3.4.4 Choice of the distance function . . . 27

4 Intrinsically Interpretable Model 29
4.1 Optimal Sparse Decision Tree (OSDT) 29

5 Comparisons 36
5.1 Introduction 36
5.2 F1 - Representativeness 37
5.2.1 F1: Normative definition 37
5.2.2 F1 : Implementation and results 38
5.3 F2 - Structure 39
5.3.1 F2 : Normative definition 39
5.3.2 F2 : Implementation and Results 40
5.4 F3 - Selectivity 41
5.4.1 F3 : Normative definition 41
5.4.2 F3 : Implementation and Results 41
5.5 F4 - Contrastivity 42
5.5.1 F4 : Normative definition 42
5.5.2 F4 : Implementation and Results 43
5.6 F5 - Interactivity 44
5.6.1 F5 : Normative definition 44
5.6.2 F5 : Implementation and Results 44
5.7 F6 - Fidelity 45
5.7.1 F6 : Normative definition 45
5.7.2 F6 : Implementation and Results 46
5.8 F7 - Faithfulness 47
5.8.1 F7 : Normative definition 47
5.8.2 F7 : Implementation and Results 48
5.9 F8 - Truthfulness 51
5.9.1 F8 : Normative definition 51
5.9.2 F8 : Implementation and Results 52
5.10 F9 - Stability 53
5.10.1 F9 : Normative definition 53
5.10.2 F9 : Implementation and Results 54
5.11 F10 - (Un)Certainty 56
5.11.1 F10 : Normative definition 56
5.11.2 F10 : Implementation and Results 57
5.12 F11 - Speed 58
5.12.1 F11 : Normative definition 58
5.12.2 F11 : Implementation and Results 58

6 Results and Conclusion 59
6.1 Results 59
6.2 Discussion 60
6.3 Conclusion 61

Chapter 1

Introduction

1.1 Contextualization

This Master's thesis focuses on eXplainable Artificial Intelligence (XAI) and aims to compare several explainability methods within the framework of a multi-class prediction task applied to an almost fully categorical dataset. The dataset consists of a survey on the habits of Quebec residents conducted in 2022. One of the primary objectives of collecting this dataset was to enable the development of a tool allowing individuals to position themselves on the political spectrum solely based on their lifestyle and demographic features (e.g., age, income). This represents a challenging and relatively uncommon case on which to test both the explainability methods and the comparison benchmark on a real-world dataset.

A central challenge of this task lies in defining criteria that allow for a meaningful comparison between fundamentally different explainability techniques. This difficulty arises from several factors, including diversity in explanation techniques, variability in explanation outputs, dependence on data type, and the lack of consensus on what constitutes a good explanation in the Machine Learning (ML) field. To address this challenge, a structured comparison framework is proposed in Section 5, largely inspired by the functionally grounded evaluation approaches described in [14] and [35]. Those scientific articles provide answers to the question of how to evaluate and compare XAI methods in a systematic way, based on the Ethics Guidelines for Trustworthy AI published by the European Union's High Level Expert Group (AI HLEG) [20].

This work begins by introducing XAI and the accuracy and interpretability tradeoff in section 1.2. Section 2 presents the datasets and the predictive model, a Neural Network (NN), used as a black-box model to be explained. Section 3 defines the theoretical foundation of the three post-hoc explainability techniques used — SHAP, LIME and Counterfactuals — while Section 4 focuses on the intrinsically interpretable model used — Optimal Sparse Decision Tree (OSDT). Those techniques were chosen because they suit the nature of the data and the task at hand. Moreover, the three post-hoc techniques are commonly used and I was interested in understanding them more deeply and to see how they compare to an intrinsically interpretable model and to each other's. The section 5 defines the characteristics providing the formal definition of a XAI's method desired properties, on which we are going to compare the 4 approaches. The section 6 discusses the findings and concludes the thesis.

1.2 Organization

1.2.1 Defintion of XAI

Following DARPA (Defense Advanced Research Projects Agency), one of the earliest actor of the field, XAI can be defined at aiming "to produce more explainable models, while maintaining a high level of learning performance (prediction accuracy) [...].”. This field resurfaced in response to the prohiminent use of Neural Network (and DNN), the need to explain those models and the increasingly more complex datasets. This is still an ongoing and recent research field without a well established consensus about the taxonomy. The term “explainable” and “interpretable” are either used interchangeably or defined as two different concepts. In the following work, one will use them interchangeably but the term “interpretable/interpretability” will mostly be used following the current ruling taxonomy in the scientific and ML community. [5] Attempts have been done to establish a taxonomy but one finds it hard to bend to since most online papers do not follow it yet.

The interpretability and accuracy trade-off

One common idea in the ML world is that there exists a tradeoff between interpretability and accuracy [6, 5, 22] :

- On one hand, simple models— such as linear and logistic regression (LR), decision tree (DT) — are intrinsincaly interpretable, given their construction, but are deemed less accurate and less fit to model complex real-life problems.
- On the other hand, Deep Neural Networks (DNNs) models yield a higher accuracy “at the cost” of interpretability. (By increasing the model’s complexity).

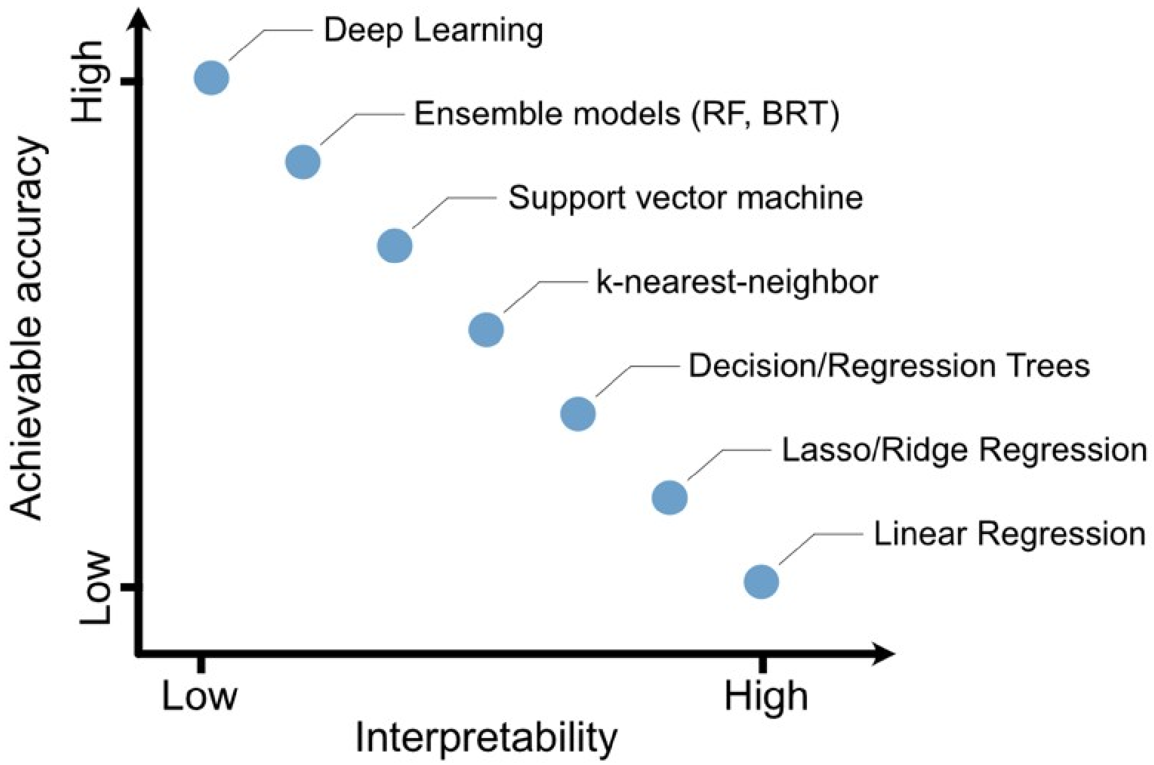

Figure 1.1: The accuracy-interpretability tradeoff : as the accuracy increases, the interpretability decreases. Source : [40]

The belief in this tradeoff partly emerged from older sources such as Leo Breiman, who introduced the Rashomon effect in statistics [13].

He asserted, more than 20 years ago, the superiority of accuracy over interpretability for statistical models.

This belief isn't simply an old idea, it is still present in the recent literature. Johansson et al.(2011) [28] write "Models exhibiting [high accuracy] are many times more complex and opaque, while interpretable models may lack the necessary accuracy." In their study of 16 biomedical classification tasks, black-box methods (like ensembles or SVMs) generally obtained higher accuracies than transparent models. This resonates with the intuition that a sufficiently complex model can fit intricate, non-linear relationships in data that a simpler model cannot capture. Those two examples might give a piece of explanation of the current inertia and tendency in the ML community ***to reach for complexity first***, motivated by the desire of maximizing accuracy in a world where data is increasingly complex.
Nonetheless, the belief into this tradeoff, or not, should not limit one to consider intrinsically interpretable models when the context allows it.

The existence of this tradeoff as general rule is debated, as it appears to be based more on intuition and heuristic than in solid empirical evidence. Most recent research tends to align with a circumstantial existence of the tradeoff. Some qualify it as a "myth" [46], while others found that when taking the user's interpretation into account that "there are contexts where black-box model can be more explainable and less confusing". [11]

The performance gap between complex and simple models exists but is not always large. Johansson et al.(2011) [28] found that, although black-box models tended to be best, the accuracy difference was typically modest — suggesting the trade-off might be a gentle slope rather than a steep cliff.

In case of tabular data, the gap in performance between interpretable and black-box model tends to be lower.[22], [46] Neural networks, in this framework, generally find no decisive advantage. It has been known for a very long time that very simple models perform surprisingly well for tabular data [25]. And this finding resurfaces in recent research.
For example, research in finance and medicine have noted that simpler models (like logistic regression or decision trees) often come within a few percentage points of the accuracy of state-of-the-art algorithms. [46]
Goethals, Martens & Evgeniou (2022) [22] analysed 90 benchmark classification tabular datasets and finds that the "trade-off exists for most (69%) of the datasets, but that somewhat surprisingly for the majority of cases it is rather small while for only a few it [the trade-off] is very large". They found that the gap in accuracy between black box and white box models is correlated to dataset's characteristics such as how difficult it is to linearly separate it. The finding reinforces our belief in the superiority of black box models in capturing the non-linearities. It also opens the door to a better identification of the conditions as to *when* using which type of models.

In the case of *raw/unstructured data*, such as images, neural networks currently have an advantage over other approaches. "The difference in performance between comprehensible models compared to a black box ones such as DL is considered "unbridgeable." Goethals, Martens & Evgeniou (2022) [22]
Yet, even in these domains, hybrid approaches are emerging (e.g. prototype-based convolutional networks [15])

"These two data extremes show that in ML, the dichotomy between the accurate black box and the less-accurate interpretable model is false" [47], or at the very least, is much more nuanced. In certain contexts, i.e. tabular data, depending on the task at hand, we must not choose accuracy *or* interpretability — we can have *both.*
Therefore, in case of tabular data, a well-constructed transparent model *could* yield both an high accuracy and be intrinsically interpretable. [48] This idea is central, to the "Interpretable Machine Learning" ("IML" in short), which one will frame as sub-field within XAI.
Those findings serve to remind the ML practitioners to explore the space of simpler models before resorting to complex black-box models if the context suits it.

The Rashomon effect

In the context of tabular data, I have previously mentioned that interpretable models can match the performance of black-box models. In principle, if one searches the space of simpler models thoroughly, one may discover a model that performs comparably to its more complex counterpart.

This observation shifts the focus of model selection: if multiple models achieve near-identical accuracy, it becomes reasonable to prioritize other properties such as interpretability or fairness. In the direct continuity of this idea, the Rashomon set refers to the ML models achieving similar performance above a given threshold. [48] Research on the Rashomon set offers theoretical support for a practice often used heuristically in applied machine learning: selecting the simplest model that performs well. For a more formal treatment of the Rashomon set and its implications, see [50].

In conclusion, this trade-off is a central idea, a knot in the field, embodied by the Post-hoc vs Intrinsically interpretable models dichotomy. The belief in a strict trade-off between accuracy and interpretability may lead practitioners to assume a forced compromise. However, empirical studies [17, 25, 48] have shown that this trade-off is often circumstantial and depends on the specific data and task at hand. The Rashomon set offers a more flexible perspective: if many models perform equally well, at least one of them may be interpretable. This insight softens the perceived trade-off and aligns with what has been observed in practice.

1.2.2 The three axis

This work will revolve around the comparison of 4 explainability approaches — three post-hoc and one intrinsically interpretable model— examined across three dimensions inspired by [6] : i) Exploratory Data Analysis, ii) Model-Level analysis and iii) Comparisons of the approaches.
The comprehension and analysis of the data will be considered as an entire part of the explainability process since there is no understanding of the model if one does not understand the task at hand.
The three axis are the following :

- (I) Exploratory Data Analysis (EDA) : the comprehension and exploration of the raw data.
- (II) Models analysis : Definition of the post-hoc interpretability methods — SHAP, LIME, Counterfactuals— and of intrinsically interpretable model —Optimal Sparse Decision Tree (OSDT).
- (III) Comparison between the methods based on functionally-grounded evaluation, i.e. quantitative metrics independent of explainees.