Why Explain a decision made by AI? Explainable AI (XAI)

Rodrigo Almeida

Abstract—In this paper, I will review the literature relating to transparency for decisions made by artificial intelligence (AI). As AI advances, not only our dependence on AI models are increasing but also a need for understanding how such results and decisions were made by the model. Due to the complexity of AI, Explainable Artificial Intelligence (XAI) has become a great choice, bringing more clarity, where outputs can be explained reducing bias and bringing transparency to the decision-making process, as opposed to "Black Box" algorithms that we cannot explain why or how the AI arrived in such a conclusion/classification/decision, it outputs directly from data, and not even who developed such algorithms are able to explain how the model arrived in such result.

1 Introduction

Explainable AI is emerging, and with it, the concerns from scientists, the general public and policymakers are also on the rise. As artificial intelligence, machine learning and deep learning are becoming integrated into applications, decisions from such models can carry consequences for a person and society, changing the focus from accuracy to the necessity of being able to explain and clarify how these decisions were made.[2] As we are becoming more dependent and relying hugely on intelligent machines, whether it is email filters, AI helping with medical diagnoses or self-driving cars, machine learning is being used across almost all sectors. Such growth in these technologies only highlights the importance of understanding the interpretability of the outcome generated by it. This is the main motivation for explainable artificial intelligence.[5] The impact of the outcome will vary based on the field. A wrong prediction for computing and business might lead to misleading recommendations, potentially affecting the revenue of such businesses, however wrong predictions in critical sectors like healthcare could put human life at risk. Opening the "black box" is incredibly important, not just for social acceptance but also for regulatory purposes. Understanding how AI systems reach their conclusions is crucial in ensuring accountability, transparency, and ethical use of these technologies, making it essential for both public trust and regulatory compliance.[2]

In Figure 1 we can easily understand the main differences between the "Black Box" and XAI.[9]

2 Definitions

Explainable AI (XAI) is a field focused on increasing human understanding of AI systems. Although terms like transparency, interpretability, and explainability are frequently used together, they can have different meanings, there are differences between these theories.[3]

- Transparency: a model deemed transparent is one that can be understood easily on its own. It is the opposite of a "black box", indicating that the inner workings of the model are understandable[1].

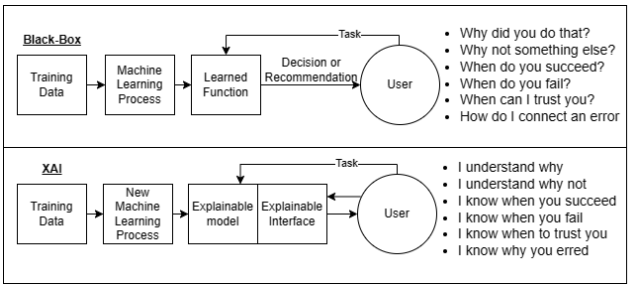

Fig. 1: XAI Concept

- Interpretability: It is the ability to explain concepts that are complex or result in a way that humans will easily understand[6].
- Explainability: it is the concept that provides explanations as a bridge between AI systems and humans. It involves the ability to create AI systems that are not only accurate but also people can understand.[6].

3 Overview of Explainable AI Algorithms

Researchers have developed many algorithms to explain AI systems. These explanations could be categorized into two primary groups: self-interpretable models refer to the algorithm model itself or a representation of the algorithm that can be directly read and comprehended by a human. In this case, the model itself serves as the explanation. On the other hand, post-hoc explanations are descriptions, explanations, or models of the algorithm often produced by separate software tools. These tools aim to provide an understanding of how the algorithm operates. Post-hoc explanations are particularly useful for algorithms for which the inner workings are not fully transparent, as they can be employed to generate insights without requiring deep knowledge of the algorithm's internal mechanisms. Instead, they rely on querying the algorithm for outputs based on selected inputs.[10].

3.1 Self-interpratable Models

Models that are self interpratable are those that serve as the explanation themselves. They not only describe the entire model but also it goes through each input, replicating the input on the self-interpretable model can providing a justification for each decision. Common self-interpretable models are decision trees and regression models. Ongoing research aims to create more interpretable models that surpass the accuracy of basic decision trees and regression models. These newer models include decision lists, decision sets, prototypes (representative samples of each class), feature combination rules, Bayesian Rule Lists, additive decision trees, and improved versions of decision trees.[10] Some sources suggest there is a trade-off between accuracy and interpretability with self-interpretable models being less accurate than post-hoc models. The challenge, in this case, is really to balance the precision of the model with what it means to humans. However, scholars like Rudin[8] and Radin[7] in their research called "Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition", disagree, explaining that there is not necessarily a trade-off between accuracy and interpretability. In many cases, interpretable models can be utilized without sacrificing decision accuracy.[10]

3.2 Pos-hoc Explanations

Post-hoc explanations refer to an explanation given after a model has been trained. In other words, the explanations are generated after the model has made the decision or prediction and not during the training phase. It is divided into two: local explanation and global explanation.

3.3 Local Explanation

Local explanations explain a subset of inputs. It focuses on explaining the output for a specific data point. The most common type is a per-decision or single-decision explanation, which gives an explanation for the output or decision on a single input.[10] These explanations focus on a single prediction and aim to answer the following question: "Why did this model make that particular prediction for that specific task?" There are some techniques used for local explanation. Some of the techniques used for local explanation are:

3.3.1 LIME - Local Interpretable Model-agnostic Explainer

LIME takes a specific decision made by an ML model, and examines nearby data points, creating a simplified and interpretable model that represents a decision made locally[10]. The default model is logistic regression. When dealing with images, LIME breaks down the image into smaller regions called superpixels. It then explores combinations of these superpixels omitting some and replacing some with black. By doing this, LIME aims to understand and explain how the model's decision is influenced by specific parts of the image.

3.3.2 SHAP - SHapley Additive exPlanations

SHAP is based on concepts of game theory and can explain the predictions by calculating the contribution of every feature to the prediction. It provides an understanding of how each feature impacts every model's prediction by considering their interaction in a consistent and fair way. This helps increase the interpretability by explaining the logic behind every prediction.[11]

3.4 Global Explanations

Global explanations produce post-hoc explanations throughout the whole dataset. It outlines the process used for the decision-making, mentioning tendencies,
features and possible biases that the model should have learned from the data. Global explanations are essential to understanding the model's behaviours and, ensuring fairness, identifying biases, and increasing trust in AI systems.
Such context in XAI helps stakeholders to cope with potential issues understanding the impact and implications of decisions made by a model. Some of the techniques used for global explanation are:

3.4.1 PDPs - Partial Dependence Plots

PDPs show the change in the response when a value changes in the feature, showing some insights about the relationship between the feature and the response. It is particularly useful when trying to determine if the relationship between a feature and the response is linear or more complex, understanding how the model behaves for individual features.[10]

3.4.2 ICE - Individual Conditional Expectation Curves

Individual conditional expectation curves are a more user-friendly way to understand how a feature influences a prediction for a single instance. by doing so, it makes it easier to understand how the model behaves for a specific case.

3.4.3 TCAV - Testing with Concept Activation Vectors

This is a global algorithm designed to explain neural networks in a way that is easier to understand. It represents the neural network state using a linear interpretation of the internals of deep learning.

3.4.4 CAVs - Concept Activation Vectors

Concept Activation Vectors are user-friendly concepts used by TCAV to explain neural networks. It is the numerical representation of a concept, making it more understandable to humans.

3.4.5 Decision sets

Opposite to black-box models, the decision sets capture the decision-making process by generating a set of rules outlining the conditions in which the model predicts certain outcomes, helping understand the model's decision boundaries.

4 Interpreting Predictions - A Case Study with the "Adult Income" dataset

As a case study, I will use the "Adult Income" data set[4] to predict whether an individual earns more than €50,000/year based on 14 features. This dataset was extracted by Barry Becker and is often cited in machine learning literature and research papers. Table 2 shows the 14 features listed by name, role type and demographic.

Aspect	Local Explanation	Global Explanation
Scope	It explains individual predictions.	It will focus on understanding the model's behaviour across the entire dataset.
Granularity	Aims to explain why a particular decision was made.	Higher level overview of the model's behaviour, bringing up general trends and feature importance across the dataset.
Techniques	LIME, SHAP, individual instance inspection.	PDPs, ICE, SP-LIME, TCAV, decision sets, summary of counterfactual rules.[10]
Use Cases	Focus on explaining why a specific prediction was made, especially in critical applications like healthcare or finance.	Focus on identifying biases, ensuring fairness, and gaining an overall understanding of the model's behaviour for regulatory compliance and model improvement.
Interpretability	Easier to understand for individual cases, focusing on the explanation specific.	Wider view of the model, making it valuable for stakeholders and policymakers seeking a general understanding of the AI system.

TABLE 1: Comparison of Local and Global Explanations in XAI

5 Interpreting Predictions - A Case Study with the "Adult Income" dataset

As a case study, I will use the "Adult Income" data set[4] to predict whether an individual earns more than €50,000/year based on 14 features. This dataset was extracted by Barry Becker and is often cited in machine learning literature and research papers. Table 2 shows the 14 features listed by name, role type and demographic.

- **Age:** It represents the age of a person, it is an important factor in predicting income because the age can reflect an individual's level of experience and earning potential.
- **Workclass:** It can be categorized in Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked. It refers to the type of employment. Status of employment is also an important factor, as income levels often vary according to employment status.
- **Fnlwgt:** Final weight represents the number of people the census believes the entry represents.
- **Education Level:** Represents the highest level of education an individual has obtained. It is divided into Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool. The education level is a key predictor of income since individuals with higher education often have access to better-paying job opportunities.
- **Education Num:** This is a numerical representation of the educational level. This also helps understand the education obtained by an individual.
- **Marital Status:** This describes the marital status of the individual. It is divided into Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, and Married-AF-spouse.
- **Occupation:** Specifies the type of education the individual is involved with. It is divided into Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces. Occupation is also a key factor which influences income, as the income levels are different depending on the occupation.
- **Relationship:** This describes the relationship status of an individual which can have an impact on household income and financial stability. It is divided into Wife, Own-child, Husband, Not-in-family, Other-relative, and Unmarried.
- **Race:** Indicates the race of the individual. It shouldn't directly affect income but could be linked to other socioeconomic factors that can impact earnings. It is divided into White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, and Black.
- **Sex:** Specifies the gender of the individual. It can influence income for reasons such as wage gaps and job segregation. It is divided into Female and Male.
- **Capital Gain:** Represents capital gains for the indi-

TABLE 2: Variable Descriptions

Variable Name	Role	Type	Demographic
age	Feature	Integer	Age
workclass	Feature	Categorical	Income
fnlwgt	Feature	Integer	—
education	Feature	Categorical	Education Level
education-num	Feature	Integer	Education Level
marital-status	Feature	Categorical	Other
occupation	Feature	Categorical	Other
relationship	Feature	Categorical	Other
race	Feature	Categorical	Race
sex	Feature	Binary	Sex
capital-gain	Feature	Integer	—
capital-loss	Feature	Integer	—
hours-per-week	Feature	Integer	—
native-country	Feature	Categorical	Other
income	Target	Binary	Income

vidual. It contributes to overall income.

- **Capital Loss:** Represents capital losses for the individual. It can impact the individual's financial situation.
- **Hours Per Week:** Indicates the number of working hours per week. The number of hours worked in a week influences the income directly.
- **Native Country:** Indicates the individual's country of origin. Depending on the country of origin the individual may be impacted by less or more opportunities. It is divided into: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad & Tobago, Peru, Hong, Holand-Netherlands.
- **Income:** This is the target variable indicating whether the individual's income exceeds $50,000 per year. This is the variable we are trying to predict.

5.1 Using Python to interpret predictions

In this section, I will use Python to interpret predictions made by a machine learning model, using the "Adult Income" described above. I will be using libraries like `pandas` for data handling, `sklearn` for building the model, and `shap` and `lime` for explainability.

5.1.1 Loading and Preparing the dataset

Using panda, the first step is to load the data and preprocess it, in a way to handle missing values and spliting the dataset into features `X` and target `y`.

5.1.2 Training a Machine Learning Model

To train the model, I will be using the `DecisionTreeClassifier` from scikit-learn for the classification task. The model will be trained using the training set and then evaluated using the test set.

```
# Train the Model
decision_tree_model = DecisionTreeClassifier(
    random_state=42)
decision_tree_model.fit(X_train, y_train)
# Make predictions
y_pred = decision_tree_model.predict(X_test)
```

Listing 1: Training a Decision Tree Classifier

5.1.3 Evaluating the Model

To evaluate the model, I will be using the `accuracy_score` from scikit-learn. The accuracy score is the fraction of predictions the model got right.

```
# Evaluate the classifier
accuracy = accuracy_score(y_test, y_pred)
print(f"Decision Tree Accuracy: {accuracy}")
```

Listing 2: Evaluating the Model

`accuracy_score` returns the accuracy of the model, which is 0.81, meaning that the model got 81% of the predictions right.

5.1.4 Global Explanation with Feature Importances

To understand the overall model behaviour, we can use the feature importances to understand which features are more important for the model. The higher the value, more important the feature is for the model.

```
# Get feature importances
importances = model.feature_importances_
# Plot feature importances
plt.barh(X.columns, importances)
plt.xlabel("Feature Importance")
plt.ylabel("Feature")
plt.show()
```

Listing 3: Feature Importances

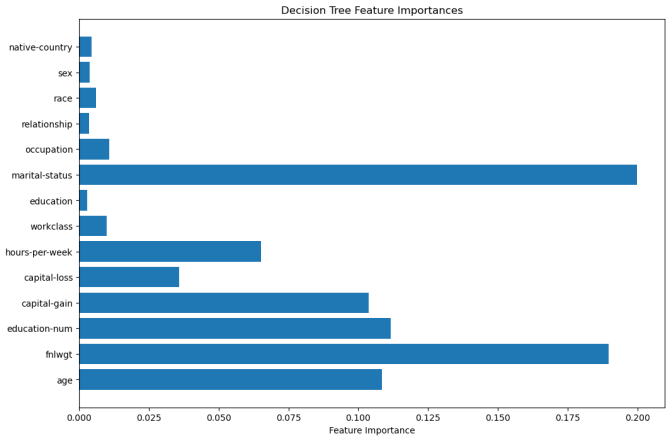

Fig. 2: Decision Tree Feature Importance

Figure 2 shows the output of code from Listing 3 feature importances for the "Adult Income" dataset. We can see that the most important features are `marital-status`, `fnlwgt`, `education-num`, `age` and `capital-gain`.

5.1.5 Local Explanation with SHAP

For local explanations, I will first use SHAP (SHapley Additive exPlanations). It provides an explanation for each feature, showing how each feature contributes to the prediction.

```
shap_values = shap.TreeExplainer(
    decision_tree_model).shap_values(X)
shap.summary_plot(shap_values, X)
```

Listing 4: SHAP Explainer

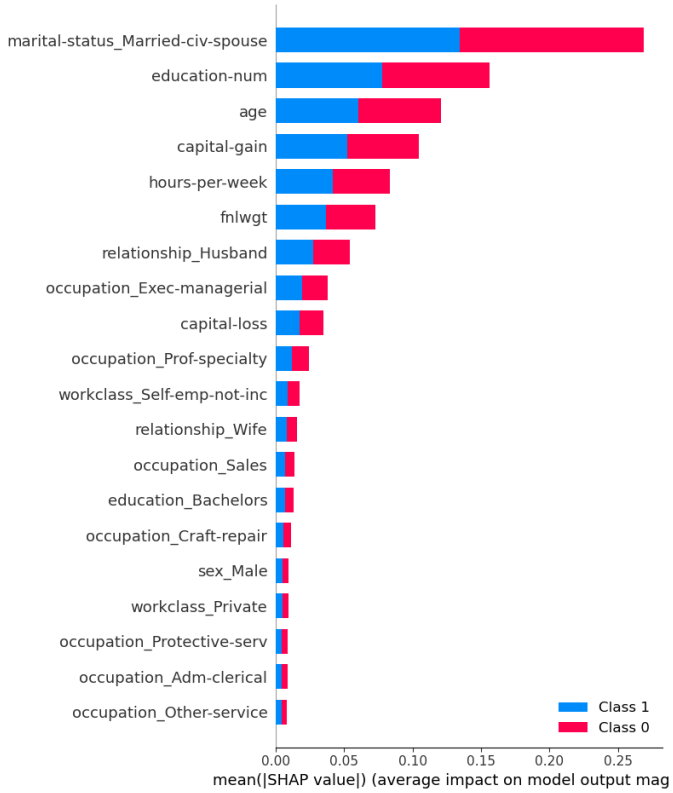

Fig. 3: SHAP Decision Plot

Note: In this SHAP summary plot, 'class 0' corresponds to incomes '≤€50,000', and 'class 1' corresponds to incomes '>€50,000'.

Figure 3 shows the output of code from Listing 4 for the "Adult Income" dataset. It demonstrates the key features influencing the model's predictions for the higher income bracket (class 1). The `marital-status` is showing as the feature with the highest mean SHAP value of 0.1345, suggesting a strong association between being married and predicting higher income. In terms of educational background, `education-num` with a mean SHAP value of 0.0779 suggesting a considerable influence, meaning that the higher the education level, the higher the income. Similarly, `age` of the individual, with a mean SHAP value of 0.0603, indicating that older individuals tend to fall into the higher income category according to the model. Financial indicators like `capital-gain` with a mean SHAP of 0.0524, and `hours-per-week` with 0.0417 are also prominent. The more a individual earns from the capital gains, more likely they are to fall into the higher income category. Similarly, the more hours an individual works per week, more likely they are to fall into the higher income category. Other features like `fnlwgt`, `relationship_Husband` and various occupational roles such as `occupation_Exec-managerial`, `occupation_Prof-specialty` have smaller but significant positive SHAP values, indicating that these features also contribute to the model's predictions for the higher income category.

5.1.6 *Local Explanation with LIME*

Another option that can be used is LIME (Local Interpretable Model-agnostic Explainer). It also provides an explanation for each feature, showing how each feature contributes to the prediction.

```
# Extracting feature names and weights from LIME explanation
feature_names, weights = zip(*exp.as_list())
# Convert to list if not already
feature_names = list(feature_names)
weights = list(weights)
# Verify the data format
print("Feature Names:", feature_names)
print("Weights:", weights)
# Creating a bar plot
plt.figure(figsize=(8, 6))
sns.barplot(x=weights, y=feature_names, palette='viridis')
plt.title('Feature Contribution to the Prediction')
plt.xlabel('Weight')
plt.ylabel('Feature')
plt.show()
```

Listing 5: LIME Explainer

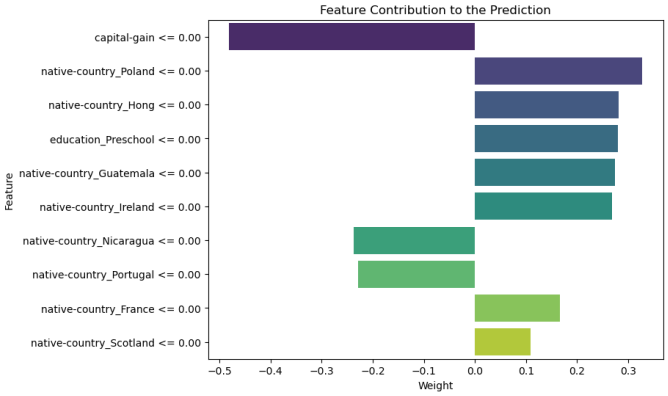

Fig. 4: LIME Decision Plot

Figure 4 shows the output of the code from Listing 5 for the "Adult Income" dataset. This output highlights the top ten features that significantly influence an individual's income category according to the model's prediction. Each feature is listed alongside a numerical weight, which quantifies the feature's influence on the prediction. The feature `capital-gain ≤ 0.00` has the most significant negative weight (-0.4817176518811231), suggesting that a strong link between having no or minimal capital gain and earning an income that's is less than €50,000 anually. On the other hand, features such as `native-country_Poland ≤ 0.00`, `native-country_Hong ≤ 0.00`, and `education_Preschool ≤ 0.00` have positive weights (0.3280648148846505, 0.2812235627400507, and 0.27962517534951975, respectively), suggesting that the absence of these conditions (e.g., not being from Poland, not being from Hong, not having preschool education) tends to associate with earning a higher income. This LIME analysis provides a localized interpretation for a specific instance, offering insights into why the model might have predicted a particular income category for an individual based on their feature values. Other features, such as `native-country_Guatemala ≤ 0.00` and `native-country_Ireland ≤ 0.00`,

also positively influence predictions towards a higher income, but their impact is somewhat less when compared to the other features. The negative weights of `native-country_Nicaragua` $\leq$ `0.00` and `native-country_Portugal` $\leq$ `0.00` indicate a tendency towards predicting lower income levels for individuals from these countries.

6 Conclusion

References

[1] Amina Adadi and Mohammed Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6 (2018), pp. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052.

[2] P. Angelov et al. "Explainable artificial intelligence: an analytical review". In: *WIREs Data Mining and Knowledge Discovery* 11 (5 2021). DOI: 10.1002/widm.1424.

[3] Anna Markella Antoniadi et al. "Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review". In: *Applied Sciences* 11.11 (2021), p. 5088.

[4] Barry Becker and Ronny Kohavi. *Adult*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20. 1996.

[5] Finale Doshi-Velez and Been Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. arXiv: 1702.08608 [stat.ML].

[6] Leilani H. Gilpin et al. *Explaining Explanations: An Overview of Interpretability of Machine Learning*. 2019. arXiv: 1806.00069 [cs.AI].

[7] Joanna Radin.

[8] Cynthia Rudin.

[9] AI concept vs. XAI Concept.

[10] P Jonathon Phillips et al. "Four principles of explainable artificial intelligence". In: *Gaithersburg, Maryland* 18 (2020).

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1135–1144. ISBN: 9781450342322. DOI: 10.1145/2939672.2939778. URL: https://doi.org/10.1145/2939672.2939778.