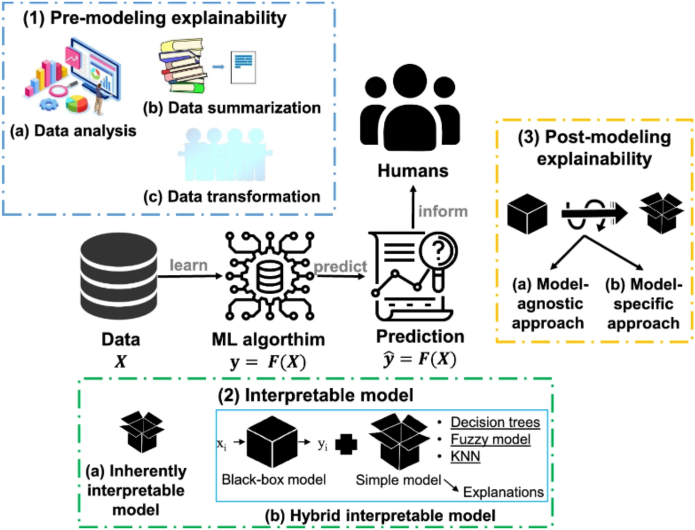
Levels of Explainability



*Explainability in XAI

Global Explanations vs. Local Explanations

Global Explanations:

- How does the model make predictions in general?
- What features influence the model's decisions the most?
- Some techniques for global explanations include: Feature importance, Decision trees, Model visualization tools like SHAP summary plots.

Local Explanations:

- Useful when someone wants to know why their application was rejected or why a particular medical diagnosis was given.
- Techniques like LIME and SHAP help us break down individual predictions.

Model-Agnostic and Model-Specific Approaches

- Suppose that the ML algorithms did not satisfy any standards to consider them an interpretable model.
- A group of approaches referred to as **post-modeling explainability** can be proposed to enable their explainability.
- The **model-agnostic approach** was devised to be implemented on any ML algorithms except for the family of deep learning models.
- The **model-specific approach** aims at addressing the explainability and interpretability for deep learning, such as CNN, RNN, and hybrid models.

Explainable AI Approaches

Feature Importance:

- Each input feature contributes differently to the model's predictions.
- Some features are more influential than others.
- Techniques rank features based on their importance.
- Helps identify key factors affecting the model's decisions.

Attribution:

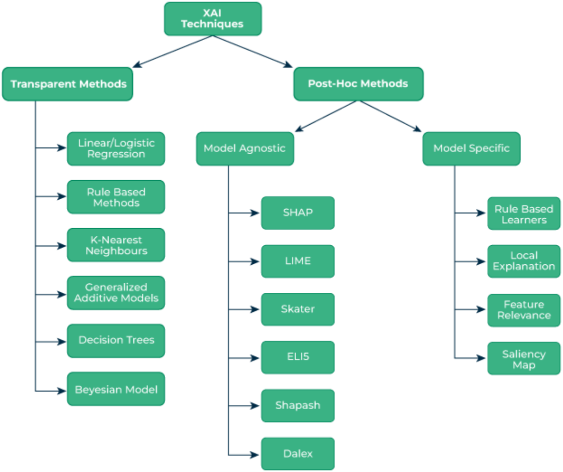
- Measures and quantifies each feature's contribution to predictions.
- Provides insight into how the model arrives at its decision.
- Techniques like SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) are commonly used.

Visualization:

- Graphical representations help in model interpretation.
- Visualization tools can display model structure, parameters, and predictions.
- Techniques include heatmaps, decision trees, and SHAP summary plots.

Explainable AI in Python

There are several approaches that you can use to implement XAI in Python, including LIME, SHAP, and ELI5.

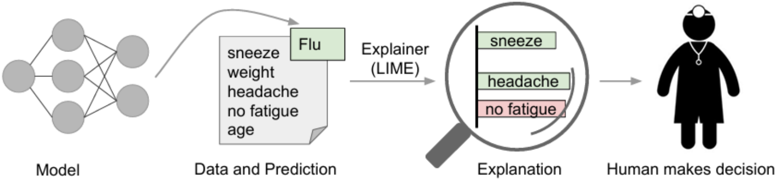


Future of Trust in XAI

- Interpretability tools catalyze the adoption of machine learning.
- It is much easier to automate interpretability when it is decoupled from the underlying machine learning model.
- Machine learning will be automated, and with it, interpretability.
- We do not analyze data, we analyze models.
- Data scientists will automate themselves.

LIME: Local Interpretable Model-Agnostic Explanations

- **Local:** Explains why a single data point was classified as a specific class.
- **Model-agnostic:** Treats the model as a black box and does not need to know how it makes predictions.



figurePaper "Why should I trust you? Explaining the predictions of any classifier" Marco Tuilo Ribeiro, Sameer Singh, Carlos Guestrin

LIME - How Does it Work

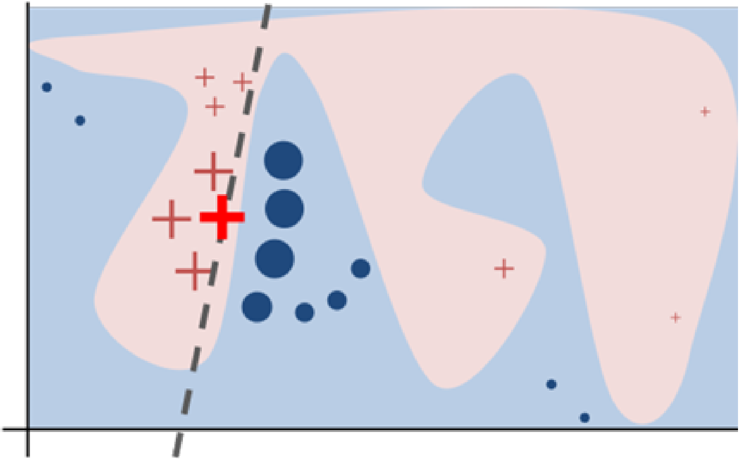


figure *Why Should I Trust You?* Explaining the Predictions of Any Classifier
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

Steps Involved in LIME's Working

1. Choose an observation to explain.
2. Create a new dataset around the observation by sampling from the distribution learned on training data.
3. Calculate distances between new points and the observation—this serves as our measure of similarity.
4. Use the model to predict the class of the new points.
5. Identify the subset of m features that has the strongest relationship with the target class.
6. Fit a linear model on the generated data in m dimensions, weighted by similarity.
7. Use the weights of the linear model as an explanation for the decision.

LIME - Advantages

- Explanations are short and contrastive. And because of human friendly explanations LIME is more suited for applications where the recipient is a lay man. However it is not sufficient for complelte attributions.
- LIME is one of the few methods that work for tabular data, text and images.
- LIME is implemented in Python (lime library) and R (lime package and iml package) and is very easy to use.

LIME - Drawbacks

- Depends on the random sampling of new points, so it can be unstable.
- Fit of linear model can be inaccurate. But we can check the r-squared score to know if that's the case.
- Relatively slow for a single observation, in particular with images.

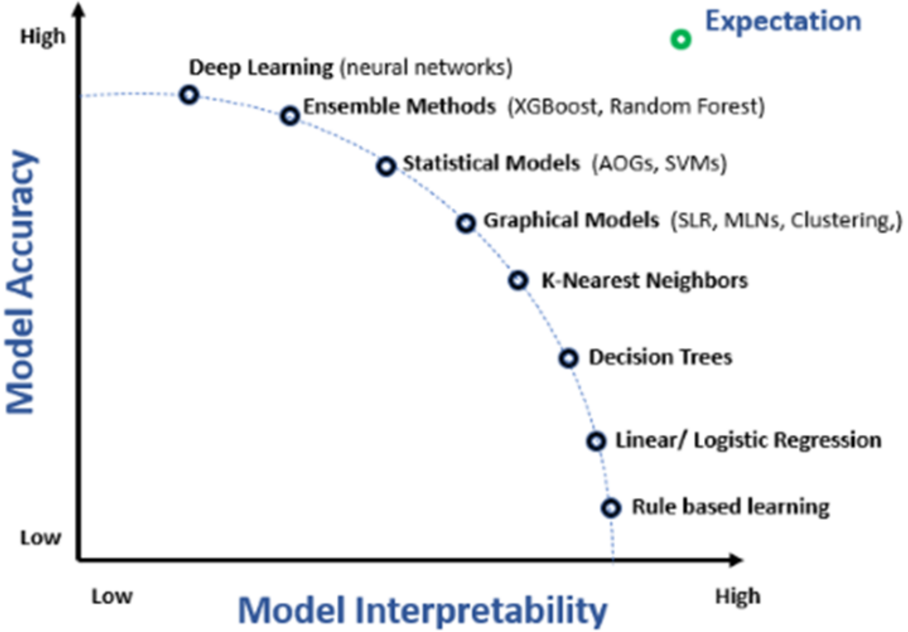
LIME-Available "Explainers"

LIME - Available Explainers
Lime supports many types of data:

- Tabular Explainer
- Recurrent Tabular Explainer
- Image Explainer
- Text Explainer

Challenges in Explainable AI (XAI)

- **Trade-off Between Performance and Explainability**
  - Deep learning models (black-box) are highly accurate but lack interpretability.
  - Simpler models are more explainable but may lose performance.
- **Evaluating Explanations**
  - No standard way to measure explanation quality.
  - Subjectivity in understanding explanations.
- **Security and Adversarial Attacks**
  - XAI methods may expose vulnerabilities.
  - Attackers can manipulate explanations to deceive users.
- **Policy and Regulations**
  - AI regulations (e.g., GDPR) require explainability, but defining "good" explanations is challenging.
  - Industries have different transparency requirements.
- **Model-Specific Explainability Issues**
  - Some XAI techniques only work for specific models.
  - A universal approach to explainability is still a research challenge.


Expectation
High
Deep Learning (neural networks)
Ensemble Methods (XGBoost, Random Forest)
Statistical Models (AOGs, SVMs)
Graphical Models (SLR, MLNs, Clustering,)
K-Nearest Neighbors
Decision Trees
Linear/ Logistic Regression
Rule based learning
Model Accuracy
Low
Low
Model Interpretability
High

Conclusion

- Explainable AI (XAI) is crucial for ensuring transparency, trust, and fairness in AI models.
- Global explanations help us understand overall model behavior, while local explanations provide insights into individual predictions.
- Techniques like SHAP and LIME improve interpretability but come with challenges, such as model-specific limitations and scalability issues.
- The rise of deep learning models has made explainability even more challenging, driving further research in XAI.
- Future advancements in XAI will focus on improving interpretability without compromising model performance.

References

- Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Available at: https://arxiv.org/pdf/1602.04938
- Explainable Artificial Intelligence: A Comprehensive Review. Available at: https://link.springer.com/article/10.1007/s10462-021-10088-y#Sec41
- O'Neil, Cathy. *"Models are opinions reflected in mathematics"*. Available at: https://everythingnewisdangerous.medium.com/models-are-opinions-reflected-in-mathematics-oneil-a93ad607f893