* **[Local Interpretability](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning" \l "Local%20Interpretability)**
  + └── [**Feature Importance**](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Feature%20Importance)
    - └── [Permutation Importance](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Permutation%20Importance)
    - ├── [SHAP Values](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#SHAP%20Values)
  + ├── [**Decision Boundaries**](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Decision%20Boundaries)
    - │ └── [LIME (Local Interpretable Model-agnostic Explanations)](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#LIME%20(Local%20Interpretable%20Model-agnostic%20Explanations))
    - ├── [Anchor Explanations](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Anchor%20Explanations)
* [**Global Interpretability**](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Global%20Interpretability)
  + └── [**Model Summarization**](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Model%20Summarization)
    - └── [Partial Dependence Plots](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Partial%20Dependence%20Plots)
    - ├── [Accumulated Local Effects (ALE) Plots](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Accumulated%20Local%20Effects%20(ALE)%20Plots)
  + ├── [**Feature Interaction**](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Feature%20Interaction)
    - │ └── [Interaction Effects](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Interaction%20Effects)
    - ├── [Feature Contribution](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Feature%20Contribution)
* [**Post-hoc Interpretability**](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Post-hoc%20Interpretability)
  + └── [**Model Agnostic**](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Model%20Agnostic)
    - └── [LIME (Local Interpretable Model-agnostic Explanations)](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#LIME%20(Local%20Interpretable%20Model-agnostic%20Explanations))
    - ├── [SHAP Values](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#SHAP%20Values)
  + ├── [**Rule-based**](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Rule-based)
    - │ └── [Decision Trees](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Decision%20Trees)
    - ├── [Rule Lists](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Rule%20Lists)
* [**Intrinsic Interpretability**](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Intrinsic%20Interpretability)
  + └── [**Transparent Models**](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Transparent%20Models)
    - └── [Linear Regression](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Linear%20Regression)
    - ├── [Decision Trees](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Decision%20Trees)
  + ├── [**Symbolic Models**](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Symbolic%20Models)
    - │ └── [Logical Rules](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Logical%20Rules)
    - ├── [Symbolic Regression](https://explorer.globe.engineer/?q=Interpretability+in%20Machine%20Learning#Symbolic%20Regression)

**Local Interpretability**

Local interpretability methods focus on explaining individual predictions. They include:

**Feature Importance**

* **Permutation Importance**: This method assesses the importance of features by randomly permuting them and observing the impact on the model's performance
* **SHAP Values**: Shapley values are an attribution method that fairly assigns the prediction to individual features

**Decision Boundaries**

* **LIME (Local Interpretable Model-agnostic Explanations)**: LIME is a model-agnostic technique that generates local explanations for individual predictions by fitting a surrogate glassbox model around the decision space of any black box model's prediction

**Anchor Explanations**: Scoped rules (anchors) are rules that describe which feature values anchor a prediction, in the sense that they lock the prediction in place

**Global Interpretability**

Global interpretability methods help understand the inputs and their overall impact. They include:

**Model Summarization**

* **Partial Dependence Plots**: Individual conditional expectation curves are the building blocks for partial dependence plots and describe how changing a feature changes the prediction
* **Accumulated Local Effects (ALE) Plots**: These plots provide a way to visualize the average effect of a feature on the prediction.

**Feature Interaction**

* **Interaction Effects**: These methods help understand how the effects of one feature depend on the value of another feature.
* **Feature Contribution**: This involves understanding the contribution of each feature to the model's predictions.

**Post-hoc Interpretability**

Post-hoc interpretability methods are used after the model has been trained. They include:

**Model Agnostic**

* **LIME (Local Interpretable Model-agnostic Explanations)**: LIME is a method that fits a surrogate glassbox model around the decision space of any black box model's prediction
* **SHAP Values**: Shapley values are an attribution method that fairly assigns the prediction to individual features

**Rule-based**

* **Decision Trees**: Decision trees are used to explain the decision-making process of the model from top-to-bottom
* **Rule Lists**: These are used to provide a list of rules that describe the model's decision-making process.

**Intrinsic Interpretability**

Intrinsic interpretability methods focus on models that are inherently interpretable. They include:

**Transparent Models**

* **Linear Regression**: This is a transparent model that provides human-understandable explanations for predictions.
* **Decision Trees**: Decision trees are inherently interpretable models that can be used for explanation

**Symbolic Models**

* **Logical Rules**: These are symbolic models that provide transparent and human-understandable explanations for predictions.
* **Symbolic Regression**: This involves finding a symbolic expression that models the relationship between the input features and the target variable.

These methods play a crucial role in understanding and explaining the predictions of machine learning models, thereby enhancing trust and transparency in AI systems.

[christophm.git...1](https://christophm.github.io/interpretable-ml-book/local-methods.html" \t "_blank)

[kdnuggets2](https://www.kdnuggets.com/2019/12/interpretability-part-3-lime-shap.html" \t "_blank)

[interpret3](https://interpret.ml/docs/lime.html" \t "_blank)

[datacamp4](https://www.datacamp.com/tutorial/explainable-ai-understanding-and-trusting-machine-learning-models" \t "_blank)

When it comes to interpreting the predictions made by a supervised learning model such as SVM or Naive Bayes for predicting controversiality of Reddit comments based on topic modeling (LDA), sentiment analysis (VADER), and user karma, you can follow these steps to ensure interpretability:

1. **Data Exploration**: Begin by exploring your data to gain insights into the relationships between the variables. Look at the distribution of the controversiality label and examine the topics generated by LDA and the sentiments identified by VADER. Analyze the correlation between user karma and controversiality.
2. **Feature Engineering**: Create new features if necessary. For instance, you could create binary features indicating whether a comment's sentiment falls within specific ranges or categories. Additionally, consider transforming the continuous user karma variable into categorical bins for easier interpretation.
3. **Model Selection & Training**: Train multiple models using different algorithms like SVM and Naive Bayes, then compare their performance to choose the best one. Use appropriate evaluation metrics such as accuracy, precision, recall, F1 score, and AUC-ROC depending on the class balance.
4. **Global Interpretability**: To understand the global impact of features on the model, use techniques such as Partial Dependence Plots (PDPs) and Accumulated Local Effects (ALE) plots. PDPs will show the overall effect of varying a single feature while keeping others constant, whereas ALE plots illustrate the average marginal effect across all possible values. In this context, generate PDPs and ALE plots for topic distributions, sentiment scores, and user karma to see their general influence on controversiality.
5. **Local Interpretability**: Apply methods such as SHAP values and LIME to investigate individual instances. Compute SHAP values for each feature contributing to a particular prediction, providing insight into why the algorithm deemed a given comment controversial. Similarly, use LIME to fit a simple, interpretable model locally around the complex model's decision boundary, helping to reveal what drove the prediction for that specific instance.
6. **Post-Hoc Interpretability**: If needed, utilize post-hoc techniques to further clarify your chosen model. Since both SVM and Naive Bayes aren't intrinsically interpretable, apply rule-based approaches such as Decision Trees or Rule Lists to represent the underlying logic governing the model's decisions. Although not perfect, these simplified representations might aid in gaining additional insights into the model behavior.

Sure! Here are more detailed steps for ensuring interpretability when working with a supervised learning model for predicting controversiality of Reddit comments based on topic modeling (LDA), sentiment analysis (VADER), and user karma:

**Step 1: Data Exploration**

* Load and preprocess your dataset containing Reddit comments, including text cleaning and tokenization
* Perform LDA to extract relevant topics and calculate topic proportions for each comment
* Utilize VADER to determine sentiment scores for each comment
* Observe the distribution of the controversiality label in your dataset
* Visualize key word clouds or bar charts representing important words and topics related to high and low controversy levels
* Investigate the relation between user karma and controversiality via scatterplots or histograms

**Step 2: Feature Engineering**

* Based on your exploration findings, engineer new features if required; e.g., creating binary features for positive, negative, and neutral sentiment categories
* Transform continuous user karma values into categorical features based on percentiles, quartiles, or custom intervals to facilitate interpretation

**Step 3: Model Selection & Training**

* Prepare training and testing datasets by splitting your original dataset (consider cross-validation strategies for better robustness)
* Choose suitable evaluation metrics according to the class balance, i.e., accuracy, precision, recall, F1 score, and AUC-ROC
* Select and train several classification models, comparing their performances:
  + Linear Support Vector Machines (SVM) with linear kernel
  + Radial Basis Function Kernel SVM
  + Naive Bayes classifiers, e.g., Multinomial NB, Bernoulli NB, Gaussian NB
* Evaluate and compare the models using selected metrics and select the most performant one

**Step 4: Global Interpretability**

* Generate Partial Dependence Plots (PDPs) for each feature:
  + Topics: Show the impact of increasing or decreasing the proportion of certain topics on controversiality
  + Sentiment scores: Demonstrate the influence of having predominantly positive, negative, or neutral sentiments on controversiality
  + User Karma: Illustrate the change in controversiality associated with various karma groups
* Produce Accumulated Local Effects (ALE) plots for each feature to complement PDP insights, emphasizing the average marginal effect instead

**Step 5: Local Interpretability**

* Implement SHAP values calculation for individual instances:
  + Display the contributions of each feature towards making a specific controversiality prediction
  + Visualize force plots revealing interactions among features affecting the prediction
* Employ LIME to build local surrogate models surrounding the decision boundaries of the primary model:
  + Present easily understood graphical summaries of driving factors behind individual predictions

**Step 6: Post-Hoc Interpretability**

* Convert the final model output into a rule-based format like Decision Trees or Rule Lists where applicable, aiming to distill the essential elements leading to the prediction without compromising too much complexity