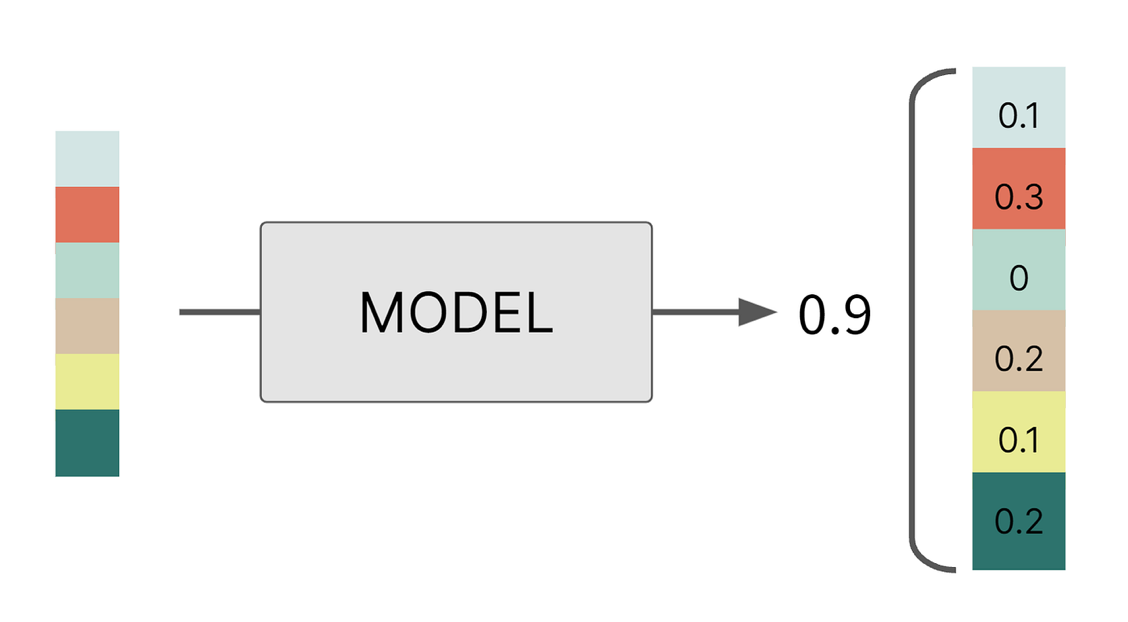
**Explainability**

Shapley Values and Global Interpretation (Python’s library for SHapley Additive exPlanations)



**Machine learning Explainability** is a critical aspect of model development and deployment, particularly as models become more complex and are used in high-stakes decision-making. Explainability can be broadly categorized into two types: local and global.

Understanding a machine learning model's decision-making process at the individual prediction level is the main goal of local Explainability. In situations like loan approval or medical diagnosis, it seeks to provide an answer to the question, "Why did the model make this particular decision for this specific instance?" The goal of global Explainability is to give light on the behavior of the model across all its predictions. "On average, which features are most important for the model's predictions?" is one of the more general (htt)issues it addresses. An overall perspective of feature importance is obtained by averaging the Shapley values over all instances in the dataset.

Choosing between local and Global Explainability depends on the specific needs of the model's stakeholders and the context in which the model is used. While local explanations are essential for individual fairness and accountability, global explanations provide oversight and understanding of the model's overall behavior, helping to identify and correct broader issues.

Advantages of Shapley Values for Global interpretation

**Fairness:** Each feature's contribution is calculated in a manner that considers all possible interactions with other features, ensuring a fair attribution of importance.

**Consistency:** If a model change such that a feature's contribution increases or stays the same regardless of other features, its Shapley value will not decrease. This consistency is crucial for reliable interpretation.

**Flexibility:** Shapley values can be applied to any model type, making them a versatile tool for global interpretation across different machine learning models.

Challenges

**Computational Complexity**: The main drawback of Shapley values is their computational expense, as calculating them requires evaluating the model's output across all possible subsets of features.

**Interpretation Complexity:** While Shapley values provide detailed insights, interpreting these values, especially in the context of complex models and interactions, can be challenging and requires a thorough understanding of the model and the domain.

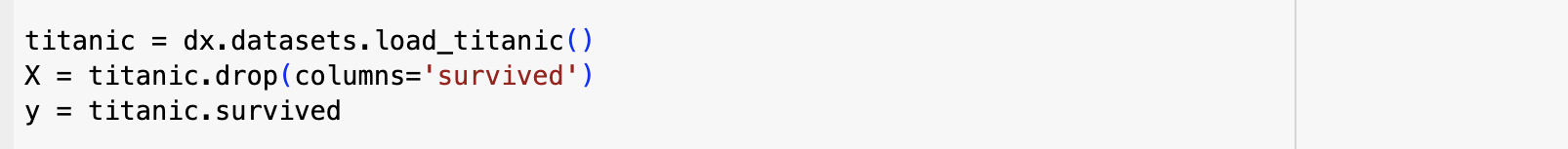
**CODE SNIPPET of working example:** shows building a predictive model using a dataset from the Titanic disaster, then explaining the prediction for a specific instance (in this case, a hypothetical passenger named Henry).

Step 1: Installing required Python libraries: ‘dalex’ for model explanations, ’pandas’ for data manipulation, and ‘sklearn’ for machine learning model building and preprocessing.

A white background with black text

Description automatically generated

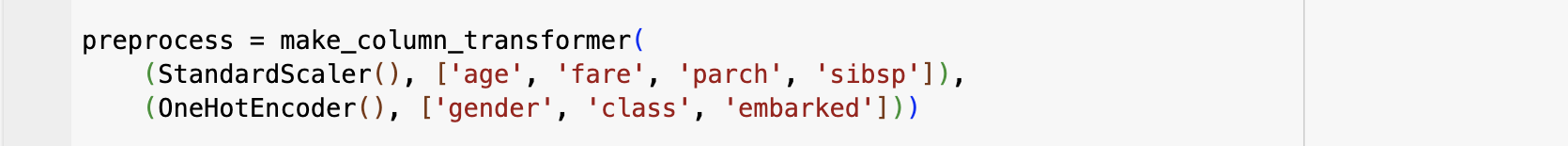
Step 2: Data Loading and Preparation: The Titanic dataset is loaded using a function from ‘dalex’. Further dividing the dataset into features(x) and variable(y). Column ‘survived’ serves as target variable for training the model.



Step3: Hypothetical condition created for prediction. A screenshot of a computer code

Description automatically generated

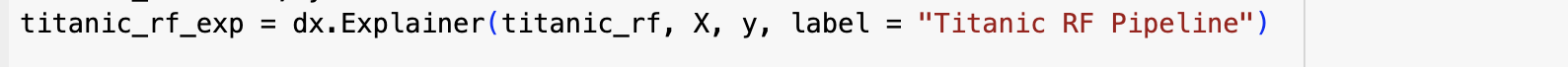
Step4: Preprocessing Pipeline: created using ‘make\_column\_transformer’. Standardization is applied for numerical columns and ‘One-hot-encoding’ for the categorical features of the dataset.



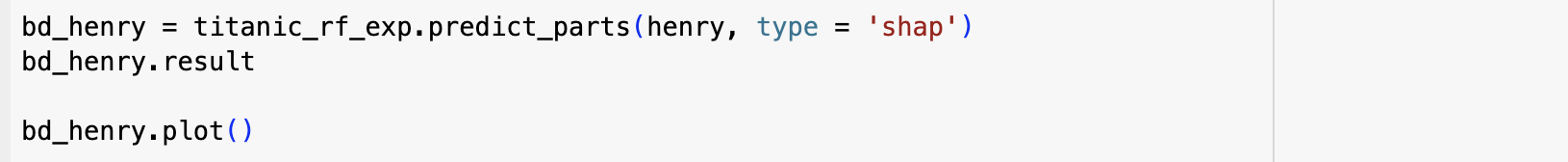
Step5: Model Training: Random Forest classifier is implemented and trained on Titanic data with pre-processing steps.



Step6: Model Explanation:



Step7: Prediction explained through visualization.



**Interpretation of the Model Output:**

A screen shot of a computer

Description automatically generated

* At first, insights on summary statistics of Explainer shows details about the data used and how well it performed the data has been used at predicting the Survival of Passenger ‘Henry’(hypothetical) at Titanic.
* ‘Predicted values’ show model’s confidence in Henry’s survival.
* Model’s predicted values on comparison with actual values produced ‘Residuals’, showing how close the predictions were.
* Shapley values are plotted to attribute the contribution of each feature to final prediction of the model. Feature ‘male’ has negative contribution, decreasing the likelihood of passenger survival whereas others tend to have positive impact.
* Being ‘male’ is most significant feature: that justifies the preference given to women and children first for survival.
* Higher socio-economic status passengers had better chances of survival. (Class= 1st and fare= 25)
* Those without families were likely to survive due to less responsibilities (sibsp =0 and parch = 0)

**Shapley Values:**

Uses: Shapley values improve machine learning model Explainability, openness, and justice, especially in sensitive areas like finance and healthcare where decision-making affects humans. This knowledge is essential for feature engineering, model optimization, resource allocation to the most important features, model debugging and validation by revealing biases or data leaks due to unexpected features. Quantifying the influence of sensitive variables on predictions helps discover and eliminate biases and promote fairness in ethical AI.

Limitations: They are difficult to compute, which takes time and resources, especially for models with many characteristics or complex structures.

1. High-dimensional data makes this challenge harder by forcing us to make estimates that may skew and reduce accuracy.
2. Shapley values are challenging to interpret since characteristics contribute and mix in complex ways, need to know a lot about the model and data.
3. Shapley values are easily changed by data or model parameters; therefore they must be regularly monitored and possibly re-calibrated, making their use even more challenging.
4. Finally, features aren't necessarily independent in practice. This makes Shapley values difficult to use directly and requires careful feature interaction consideration.

Relevance to Compliance:

1. Transparency and Explainability: With regulatory guidelines like GDPR implemented gives right to explanation to individuals affected by automated decision-making. Shapley values can illustrate how each attribute affects/contributes to model’s prediction and making decisions more understandable.
2. Fair and bias Mitigation: Shapley values can assist eliminate biases and ensure fair outcomes by determining if and how such traits affect model predictions.
3. Auditability and Accountability: They allow for systematic accounting of input feature contributions, which might be vital for regulatory examination of model behaviors and judgments.

Issues with Compliance: The general application is problematic due to their computational complexity and scalability issues, which can limit their applicability, particularly for real-time monitoring in models with many features. Shapley values provide complicated explanations that can be misunderstood, especially by non-technical stakeholders. This makes it difficult to make sure that explanations are clear and accurate. Since Shapley values must be calculated using potentially sensitive data, data privacy becomes a major risk that must be addressed by adhering to data protection rules like GDPR. Shapley values' standardization for compliance is made more difficult by the variations in regulatory requirements throughout jurisdictions, necessitating flexible methods to satisfy a range of criteria for transparency and explanation. Furthermore, people without the requisite technical expertise may be excluded due to the intricacy of the mathematical ideas behind Shapley values, which raises questions regarding information accessibility and equity.

References:

<https://ema.drwhy.ai/shapley.html#SHAPPythonCode>

<https://chat.openai.com/>

<https://towardsdatascience.com/the-shapley-value-for-ml-models-f1100bff78d1>

<https://towardsdatascience.com/understand-the-working-of-shap-based-on-shapley-values-used-in-xai-in-the-most-simple-way-d61e4947aa4e>