

Advancing ML Model Evaluation: A Comprehensive Platform for Automated Testing and Analysis For Supervised Machine Learning Tasks

Project Supervisor: Mohamed Maher

THE TEAM

- **Joanes De Miguel** - Artificial Intelligence BSc, 4th year (University of the Basque Country, EHU/UPV)
- **Lukas Arana** - Artificial Intelligence BSc, 4th year (University of the Basque Country, EHU/UPV)
- **Oier Ijurco** - Artificial Intelligence BSc, 4th year (University of the Basque Country, EHU/UPV)



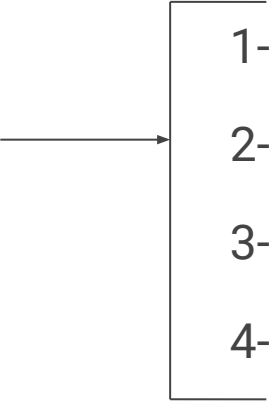
PROJECT DESCRIPTION

- Create a platform that functions as an intelligent evaluator
- The platform has a pipeline that conducts a series of tests and analysis about the dataset and performance of a specific model
- Requires a fitted model and the test split of the dataset as a minimum
- Everything is put together in a web application made with Django




APPROACH OF THE PROBLEM

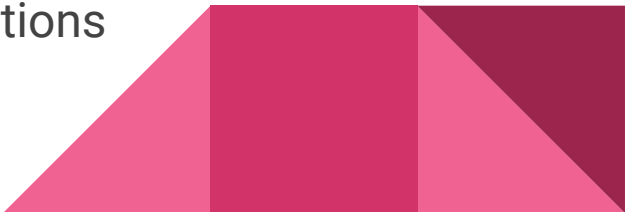
- What different evaluations do our platform need to provide?

- 
- 1- Performance Evaluation
 - 2- Inference Time and Environmental Impact
 - 3- Model Bias and Variance Analysis
 - 4- Ethical Input Feature Analysis

- How are we going to present our platform and pipeline?



Django's templating system and its framework for building web applications



TOOLS USED



SHAP



DJANGO

- It is a high-level, Python web framework designed for rapid development of websites.
- We've made a very basic website just to have an Interface for our pipeline
- Our website lets the user upload the necessary files (the pickle model, the test split...) and executes the pipeline, providing the analysis about the dataset and performance of the model



THE INPUT

- A pickle file containing the fitted model
- The test split of the dataset
 - `X_Test`
 - `y_Test`
- Whether it is a Classification or a Regression task



THE DIFFERENT EVALUATIONS

- 1- Performance Evaluation
- 2- Inference Time and Environmental Impact
- 3- Model Bias and Variance Analysis
- 4- Ethical Input Feature Analysis



1- Performance Evaluation

The platform generates a comprehensive performance report using and displaying the following metrics:

- Classification
 - Accuracy
 - AUC
 - F1-Score
 - FPR, FNR, TRP, TNR
 - Cross-Entropy Loss
 - Confusion Matrix
 - Regression
 - MAE
 - MSE
 - RMSE
 - Mean Bias Deviation
 - R-Squared Error
- 

2- Inference Time and Environmental Impact

- Measurement of the average time taken to make an inference and conversion to CO2 emission using fixed hardware specs and energy consumption conversion.
- Measurement of the model storage size in GB.
- Summary Paragraph using a hard-coded schema



3- Model Bias and Variance Analysis

Analysis of bias and variance in a model for a specific task, considering the impact of different features on model performance:

- It calculates the percentage of null values for each feature
- Assessment of bias introduced by unique values of categorical or numerical features
- Measurement of variance by perturbing feature values.



4- Ethical Input Feature Analysis

- We used ADA 002 for embeddings
- Use LLMs to understand if any of the features is unethical such as Gender, Race, Nationality, Religion, etc.
- Measure the feature importance using Global Interpretability Techniques like SHAP values



EXAMPLE OUTPUT FOR TITANIC DATASET

Classification Metrics:

Accuracy: 0.8380

AUC: 0.8220

F1 Score: 0.7883

True Positive Rate: 0.7297

True Negative Rate: 0.9143

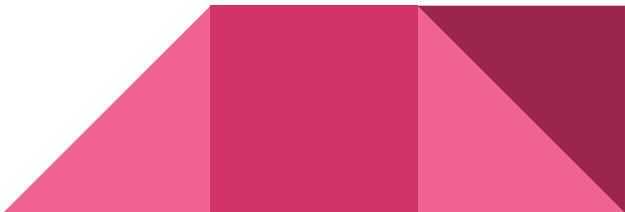
False Positive Rate: 0.0857

False Negative Rate: 0.2703

Cross entropy loss: 0.5043

Confusion Matrix:

```
[[96  9]
 [20 54]]
```



Average Inference Time: 0.0053 seconds
CO2 Emission per Prediction: 0.0003 kgCO2
CO2 Emission in total: 0.0497 kgCO2
Estimated Model Storage Size: 0.0016 GB


Our platform evaluates the model's inference time and environmental impact. The average time taken to make an inference is measured and converted to CO2 emissions, assuming Coal as a fuel for energy generation and Xeon 2.2 GHz Core. The mass of emitted CO2 by a single prediction is estimated to be approximately 0.0003 kgCO2. The estimated model storage size is 0.0016 GB.



Feature Null Values: {0: 0.0, 1: 0.0, 2: 0.0, 3: 0.0, 4: 0.0, 5: 0.0, 6: 0.0, 7: 0.0, 8: 0.0, 9: 0.0}

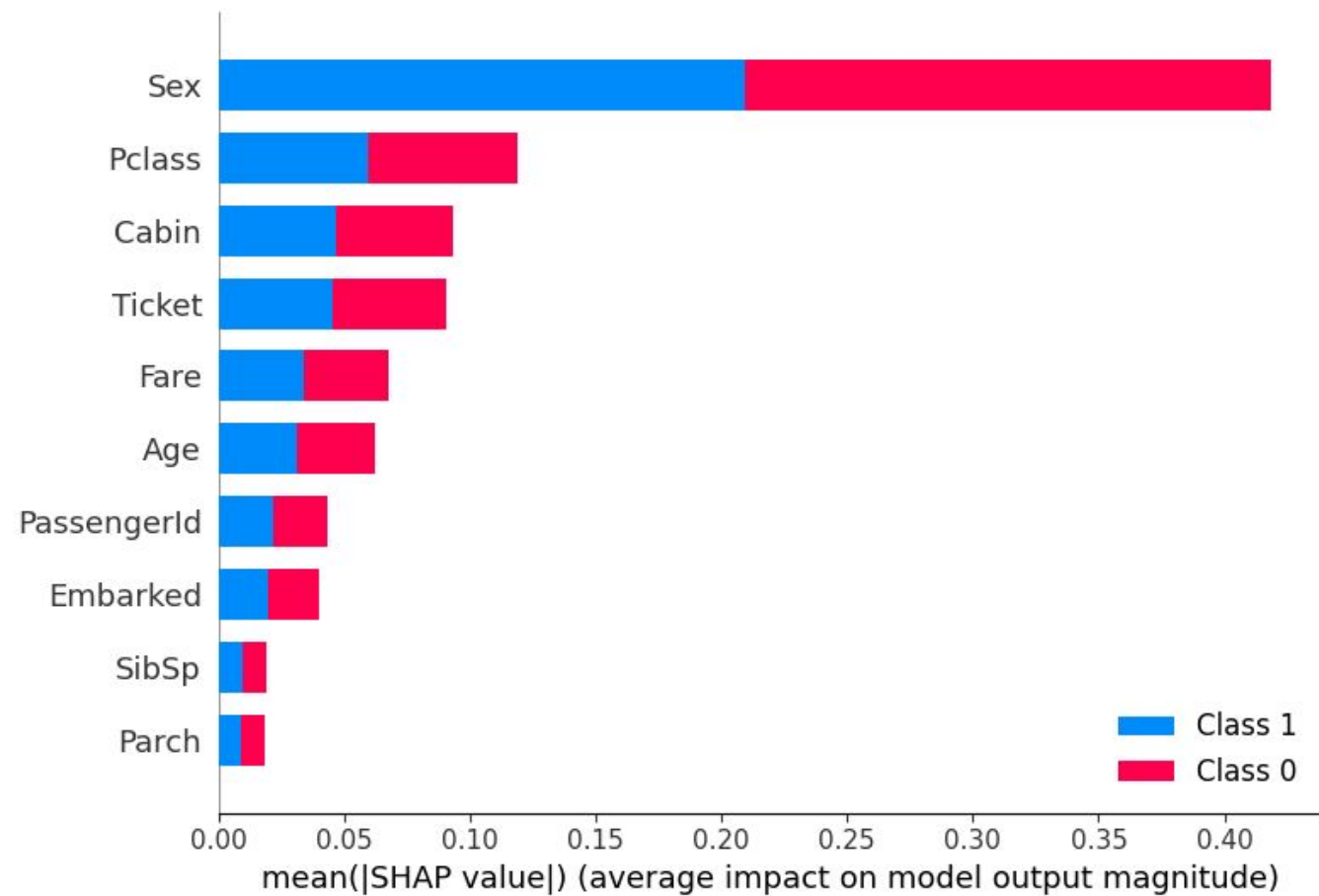
Category Bias List: [0.6000749040292133, 0.3502569738981535,
0.3632193933505568, 0.4544741616485221, 0.3341537013676138,
0.4040682765629754, 0.5952585589457091, 0.47116608612849414,
0.40339907581748263, 0.2975028649890514]

Feature Perturbation List: [0.050279329608938994, 0.07821229050279399,
0.32960893854748896, 0.07821229050279399, 0.005586592178770999,
0.011173184357541999, 0.08938547486033599, 0.050279329608938994,
0.08379888268156499, 0.039106145251396995]



```
[{'feature_description': 'PassengerId', 'is_unethical': array([[False]])},  
 {'feature_description': 'Pclass', 'is_unethical': array([[True]])},  
 {'feature_description': 'Sex', 'is_unethical': array([[True]])},  
 {'feature_description': 'Age', 'is_unethical': array([[True]])},  
 {'feature_description': 'SibSp', 'is_unethical': array([[True]])},  
 {'feature_description': 'Parch', 'is_unethical': array([[False]])},  
 {'feature_description': 'Ticket', 'is_unethical': array([[True]])},  
 {'feature_description': 'Fare', 'is_unethical': array([[False]])},  
 {'feature_description': 'Cabin', 'is_unethical': array([[False]])},  
 {'feature_description': 'Embarked', 'is_unethical': array([[False]])}]
```





FUTURE STEPS

- Improve Front-End (Django)
- Implement MLFlow
- (...)



THANK YOU FOR YOUR ATTENTION

Repository with the code: https://github.com/joanes28/second_project

