Report - Explainable AI (xAI) Dashboard for UCI Adult Dataset

1. Problem Understanding & Rationale

AI models, particularly black-box ones such as XGBoost, attain great accuracy but are not interpretable. Grasping the reasons behind a model's particular prediction is essential for:

Trust & Transparency: Users can understand the reasons behind decisions.

Equity & Morality: Personal attributes such as age, gender, and ethnicity can be tracked for prejudice.

Cybersecurity: Tracking misuse or unusual predictions enhances system resilience.

Aim: Create an engaging dashboard that illustrates the rationale behind an AI model's predictions for the UCI Adult dataset, utilizing SHAP explanations

2. Dataset Description

Dataset: UCI Adult Income Dataset

- **Samples:** 32,561
- **Features:** 14 input features + 1 target (income >50K or <=50K)
- **Feature Types:**
    - **Categorical:** workclass, education, marital-status, occupation, relationship, race, sex, native-country
    - **Numerical:** age, fnlwgt, education-num, capital-gain, capital-loss, hours-perweek

Target Variable:

- 0 → Income ≤ 50K
- 1 → Income > 50K

Notes on Indices:

- **Sample index** in the dashboard (0,1,2,…) refers to the **row in the test set**. Selecting an index shows the model prediction and SHAP explanations for that specific sample.

3. Design & Implementation Approach

Models Trained:

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. XGBoost (**Best-performing model**)

Evaluation Metrics:

- Accuracy, ROC-AUC, F1-score

Model	**Accuracy**	**ROC-AUC**	**F1-score**
XGBoost	0.87	0.93	0.71
Random Forest	0.85	0.90	0.67
Logistic Regression	0.84	0.90	0.66
Decision Tree	0.84	0.89	0.62

Explainability Approach:

- **SHAP (SHapley Additive exPlanations):**
  - Provides global feature importance across all test samples. o Gives local explanations for individual predictions.
  - Outputs human-readable statements showing how features increase/decrease prediction probability.

Note: Decision Tree and LIME explanations were implemented in the code but are **not included in the deployed prototype**, focusing on XGBoost and SHAP for simplicity and clarity.

Dashboard Features:

1. **Sample Selection:** Choose any test sample index to inspect its prediction.
2. **Model Prediction & Comparison:** Shows XGBoost prediction with probability.
3. **SHAP Explanations:**
   - Global feature importance plots.
   - Local explanation table with features, SHAP values, and impact.
4. **Trust & Safety:** Highlights sensitive features and their influence on predictions.
5. **Cybersecurity:** Optional user login system with registration to demonstrate secure access.

4. Results & Observations

Example Predictions:

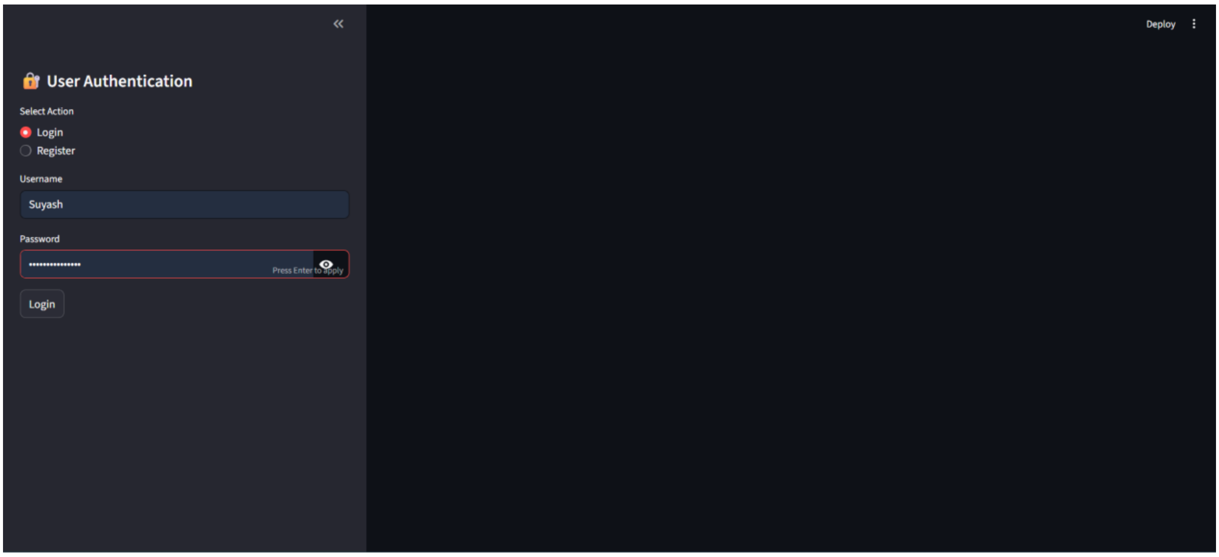
- **Sample Index 16:**
  ○ XGBoost Prediction: `0` (Income ≤50K, probability 0.99) ○ Key SHAP Influences:

Feature	**SHAP Value**	**Impact**
marital-status_Married-civ-spouse	-1.623	Decreased
age	-1.023	Decreased
education-num	0.585	Increased
sex_Female	-0.395	Decreased
hours-per-week	-0.236	Decreased
capital-gain	-0.180	Decreased
occupation_Sales	0.123	Increased

- **Interpretation:** Negative SHAP values decrease the probability of class 0, positive values increase the probability.

General Observations:

- Age and marital status significantly impact income predictions.
- Sensitive features are monitored to ensure fairness.
- Local explanations help users understand **why the model made a certain decision**.



🔐 User Authentication

Select Action

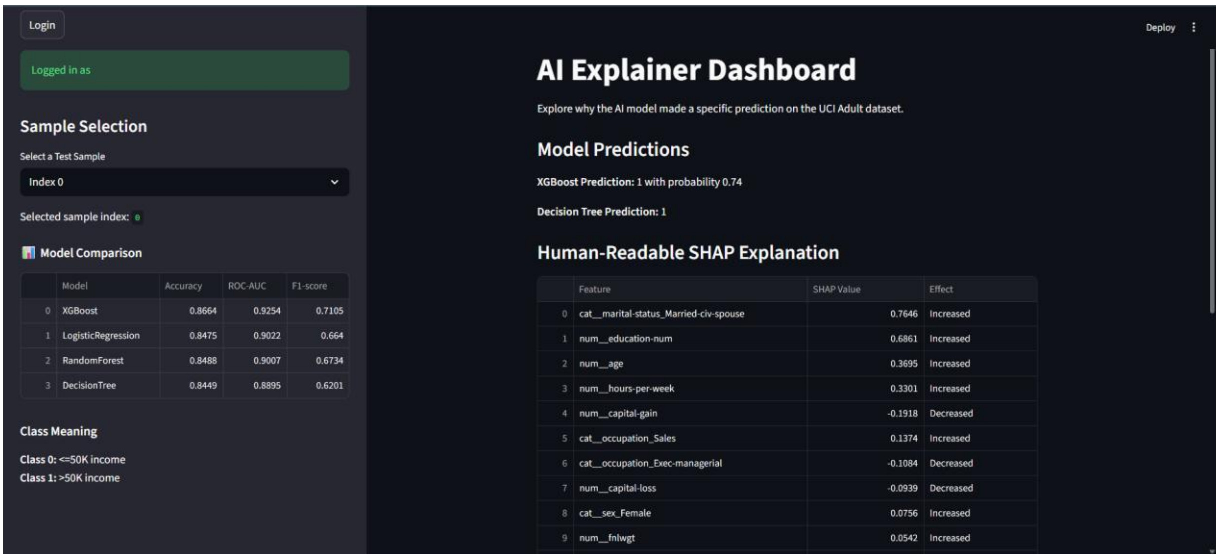
● Login
○ Register

Username

Suyash

Password

•••••••••••••••



Logged in as

Sample Selection

Select a Test Sample

Index 0

Selected sample index: 0

📊 Model Comparison

	Model	Accuracy	ROC-AUC	F1-score
0	XGBoost	0.8664	0.9254	0.7105
1	LogisticRegression	0.8475	0.9022	0.664
2	RandomForest	0.8488	0.9007	0.6734
3	DecisionTree	0.8449	0.8895	0.6201

Class Meaning

Class 0: <=50K income
Class 1: >50K income

AI Explainer Dashboard

Explore why the AI model made a specific prediction on the UCI Adult dataset.

Model Predictions

XGBoost Prediction: 1 with probability 0.74

Decision Tree Prediction: 1

Human-Readable SHAP Explanation

	Feature	SHAP Value	Effect
0	cat__marital-status_Married-civ-spouse	0.7646	Increased
1	num__education-num	0.6861	Increased
2	num__age	0.3695	Increased
3	num__hours-per-week	0.3301	Increased
4	num__capital-gain	-0.1918	Decreased
5	cat__occupation_Sales	0.1374	Increased
6	cat__occupation_Exec-managerial	-0.1084	Decreased
7	num__capital-loss	-0.0939	Decreased
8	cat__sex_Female	0.0756	Increased
9	num__fnlwgt	0.0542	Increased

5. Security, Ethical, & Governance Considerations

- **Sensitive Data:** Features like sex, race, and age are highlighted to detect potential bias.
- **User Authentication:** Only authorized users can access the dashboard (optional login system).
- **Explainability:** Using SHAP increases trust in AI decisions.
- **Audit Trails:** User actions and predictions are logged for accountability.

Ethical Implications:

- Transparent AI can prevent discriminatory outcomes.
- Users can validate model predictions before taking automated actions.

6. References

1. https://archive.ics.uci.edu/ml/datasets/adult
2. https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learninginterpretability
3. https://joblib.readthedocs.io/en/stable/