Assignment 14

Fairness and Explainability Analysis on the Titanic Dataset

1. Introduction to Ethical AI

In recent years, the importance of ethical considerations in Artificial Intelligence (AI) has grown significantly. As machine learning models are increasingly deployed in high-stakes domains such as finance, healthcare, and criminal justice, concerns regarding fairness, accountability, and transparency have come to the forefront.

Bias in AI models can lead to discrimination against certain groups, particularly those defined by sensitive attributes such as gender, age, or race. Explainability methods such as SHAP and LIME help demystify model behavior, making it possible to identify sources of bias and interpret model predictions.

This project explores fairness and explainability using the Titanic dataset. We train a logistic regression model to predict survival and evaluate both performance and fairness. Additionally, we apply interpretability tools to understand the model's decisions.

---

2. Dataset and Preprocessing

Dataset Description

We used the public Titanic dataset, which contains information about passengers aboard the Titanic, including:

- **Features**: Pclass, Sex, Age, Fare, Embarked
- **Target Variable**: Survived (1: Survived, 0: Did not survive)
- **Sensitive Attribute**: Sex (used to analyze fairness across gender)

Preprocessing Steps

- Removed missing values for Age, Sex, Fare, and Embarked.
- Encoded categorical variables:
    - Sex: male=1, female=0
    - Embarked: mapped to integer codes
- Selected input features (X) and target label (y).

- Split the data into training and testing sets (70/30 split).

---

3. Model Architecture and Evaluation

Model

- **Algorithm**: Logistic Regression
- **Library**: Scikit-learn
- **Hyperparameters**: Default (max_iter=1000 to ensure convergence)

Performance Metrics

- **Accuracy**: 78%
- **Confusion Matrix**: [[108 14]
- [33 59]]
- **Classification Report**:
    - Precision: 0.8 (positive class)
    - Recall: 0.72 (positive class)
    - F1-score: 0.76

The model performs reasonably well on the test data, but performance alone does not reveal potential bias.

4. Fairness Analysis

Using the Fairlearn library, we analyzed model behavior across the gender attribute (Sex).

Metrics Analyzed

Metric	**Female (0)**	**Male (1)**
Accuracy	0.84	0.74
Selection Rate	0.83	0.065
False Positive Rate	0.47	0.04

Metric	Female (0)	Male (1)
True Positive Rate	0.94	0.11

Observations

- **Selection Rate**: Women were predicted to survive much more often than men (0.68 vs 0.30).
- **True Positive Rate**: Model is much better at identifying actual survivors among women.
- **Fairness Concern**: The model is biased in favor of women, likely due to historical survival patterns on the Titanic.

Visualization

A bar chart showed clear disparity across all fairness metrics, highlighting a need for fairness-aware modeling techniques.

5. Explainability Analysis

SHAP (SHapley Additive exPlanations)

- **Global Summary**:
    - Most impactful features: Sex, Fare, Pclass, Age
    - Sex had the highest contribution to prediction of survival
- **Local Explanation**:
    - SHAP waterfall plots explained individual predictions, showing how each feature pushed the model's output up or down.

LIME (Local Interpretable Model-Agnostic Explanations)

- LIME was used to generate explanations for individual cases.
- It confirmed SHAP's findings, revealing that gender and ticket class heavily influenced predictions.

6. Ethical Considerations and Recommendations

Key Issues

- **Gender Bias**: Model favors female passengers due to real-world historical survival outcomes. However, this perpetuates biased outcomes.
- **Model Trust**: Without interpretability tools, it would be difficult to detect and address these issues.

Recommendations

- Use reweighting techniques or fairness-aware classifiers (e.g., adversarial debiasing).
- Conduct post-processing to balance true positive and false positive rates across groups.
- Use explainability tools regularly in ML development pipelines.

7. Conclusion

This project demonstrated the application of fairness and explainability analysis on a real-world dataset. While the logistic regression model performed well, it exhibited clear gender-based bias. Through tools like Fairlearn, SHAP, and LIME, we identified both the source and nature of the bias.

Future Improvements

- Use a larger, more balanced dataset.
- Apply fairness-aware algorithms during training.
- Regular audits using interpretability and fairness tools.