CSI 5195 ETHICS ARTIFICIAL INTELLIGENCE

AI Fairness in Employee Promotion Systems

A Subgroup Analysis of Fairness, Bias, and Explainability

Jenifer Yu Kelvin Mock Yifan Qin Shengchen Liu Jing Hu

Group 12

PROMOTION FAIRNESS IN AI

Introduction to Project

This presentation explores the critical aspects of gender fairness in AI-driven employee promotion systems, highlighting key challenges and potential solutions.

Exploring AI Fairness in Promotions

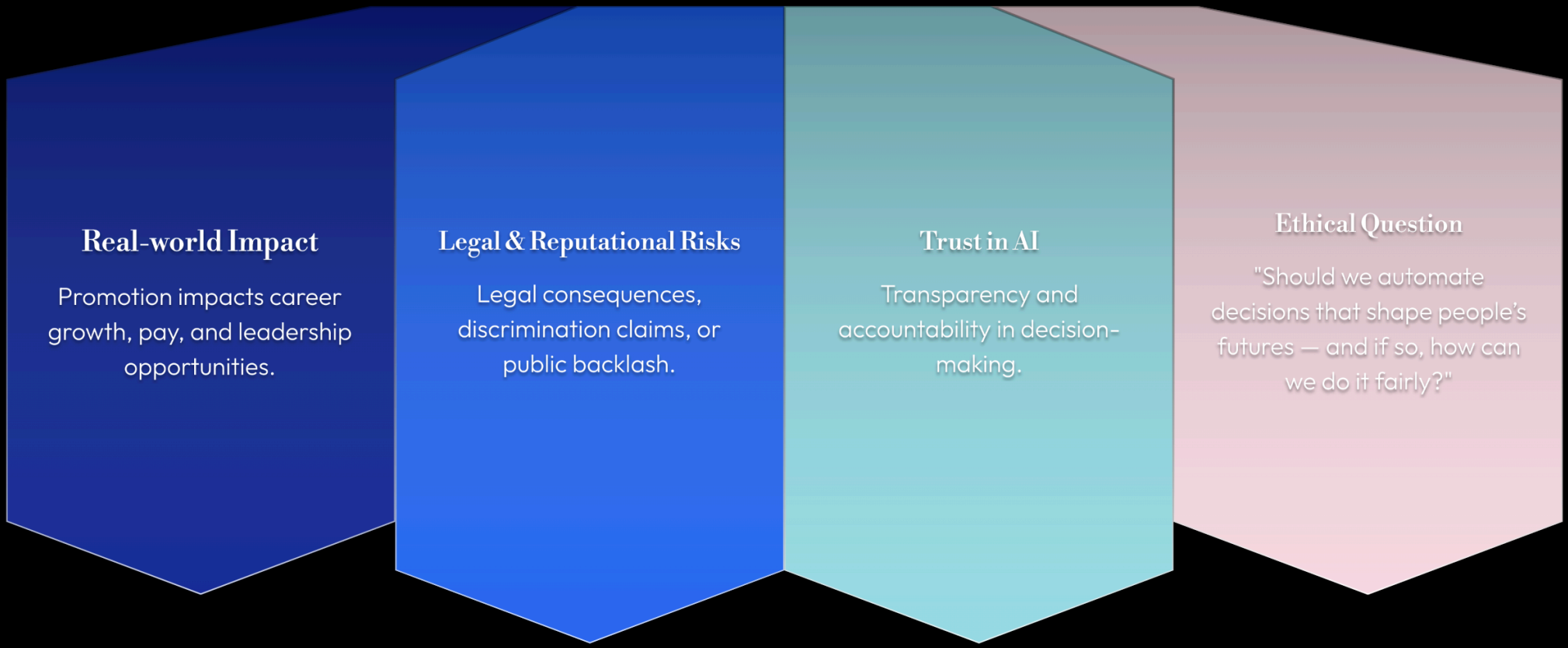
AI is increasingly used to assist promotion decisions.

Whether AI models employees fairly in promotion?

These systems often learn from biased historical data.

This can lead to gender bias — promoting men more often than women. outcomes.

Why Should We Care?



SECTION 01

Literature Review

This review examines existing literature on gender fairness within AI-powered employee promotion systems, highlighting challenges and solutions.

Defining Fairness in AI: Key Concepts and Criteria

Core Assumption: Equal treatment across different identifiable groups.

Group-Based Fairness Metrics

- Demographic Parity: Equal acceptance rates across different groups.
- Accuracy Parity: Equal prediction accuracy across subgroups.
- Predictive Rate Parity: Model predictions should align with a candidate's qualifications.

Fairness Perspectives

- Individual Fairness: Similar individuals should receive similar outcomes.
- Group Fairness: Ensure fair treatment across identifiable subgroups.
- Counterfactual Fairness: A model's decision should remain the same if a protected attribute (e.g., gender) were changed.

Beyond Metrics

- Bias Detection & Mitigation: Remains a human responsibility
- Final Decision-Making: AI should support, not replace, human judgement
- Goal: Build fair, transparent, and trustworthy AI systems

Source: Dena F.M. & Nihar R.M. (IEEE ISTAS, 2019)

Fairness Detection/Mitigation

- Pre-processing
  - Modifying the dataset – reweighting / resampling
  - Edit Features, Labels according to fairness definitions
  - Remove any obvious sensitive attributes
  - Extendable to Multi-Modal or Multi-Task Learning
- In-processing (Optimization)
  - Achieve Accuracy Parity – with explanation to a decision
  - High Performance! (but accuracy might be modified)
  - Modifying the Model itself Be careful when outsourcing!
- Post-processing (Counterfactuals)
  - Modifying the Outcome
  - Thresholding Completed Classifications
  - Make use of Explainability (XAI) tools

Explainability (XAI)

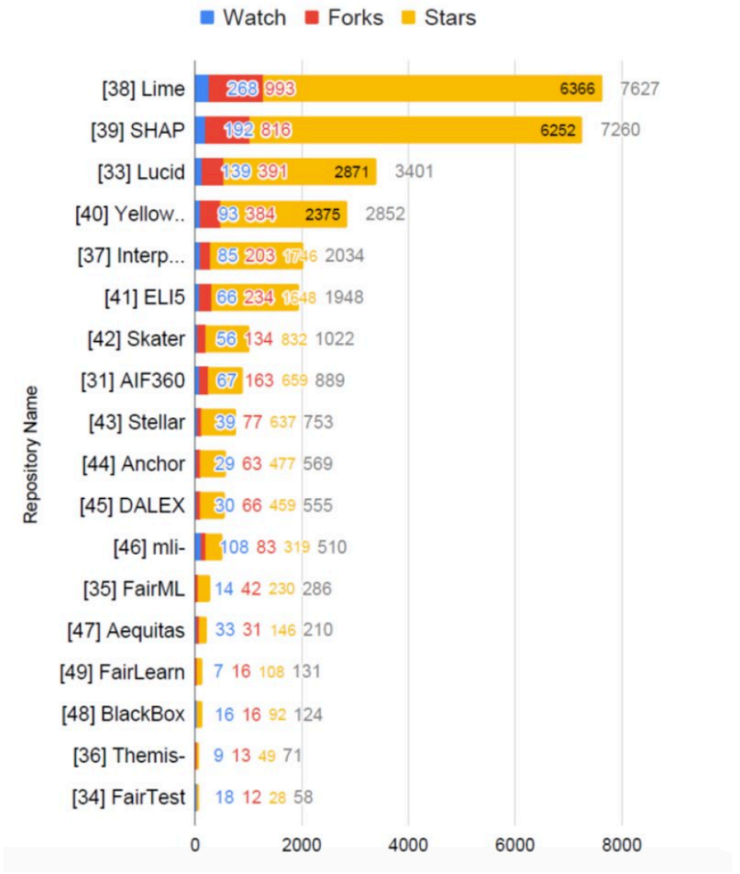
- XAI is post-processing
- Purpose: Bias detection
- Understandable Feedback
- LIME is proven the most common tool for explaining outcomes

✓ Internally improve performance

✓ Improve perceived fairness

✓ Explain counterfactuals

✓ Trustability of a black box



Comparison of Popularities of interpretability tools on GitHub

SECTION 02

Methdologies & Experiment

How do we know whether current promotions are fair?

Dataset

Original Dataset Size

54808 rows × 13 columns

After Cleaning

Reduced to 48,660 rows, ensuring higher quality for analysis.

Target Label Definition

The target label in this analysis is 'is_promoted', indicating whether an employee has been promoted or not.

Sensitive Attribute

Gender is identified as a sensitive attribute, categorized as 'm' for male and 'f' for female.

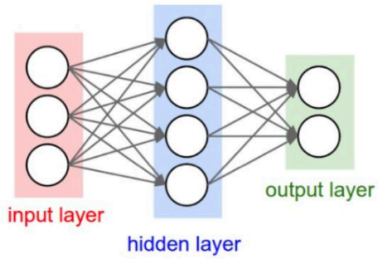
Gender	Not Promoted	Promoted	Promotion Rate (%)
Male	30983	2869	8.47%
Female	13445	1363	9.20%

Sensitive Attributes vs Target Label Distribution

Methodologies

Baseline Model

- Model: Standard MLP Classifier
- Input: All features except gender
- Evaluation: Overall Performance vs Performance across subgroups

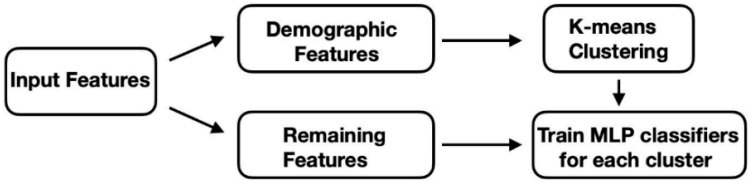


$$\mathbf{h} = f(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)})$$
$$\mathbf{z} = \mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)}$$
$$\mathbf{y} = \text{softmax}(\mathbf{z})$$

Baseline Model: No explicit fairness handling.

Cluster-Based Fairness-Aware Model

- Step 1: Demographic Clustering
  - Features used: Demographic Features (First 40 features).
  - Method: K-Means Clustering.
  - Output: k clusters (latent subgroups).
- Step 2: Cluster-specific Classification
  - Train one MLP per cluster using remaining features.
  - Predict by: Cluster Assignment → Corresponding MLP → Output.



$$P(\text{Outcome} \mid \text{Demographics}) \approx P(\text{Outcome} \mid \text{Gender})$$

New model: Clustering approximates sensitive structure without using sensitive attributes.

No gender feature used during training in both models.

Experimental Setup

Exploration of AI model training and fairness

- **Data Pre-Processing Steps**

  Included data cleansing, normalization, and an 80-20 train-test split.

- **Baseline and Fairness-Aware Models**

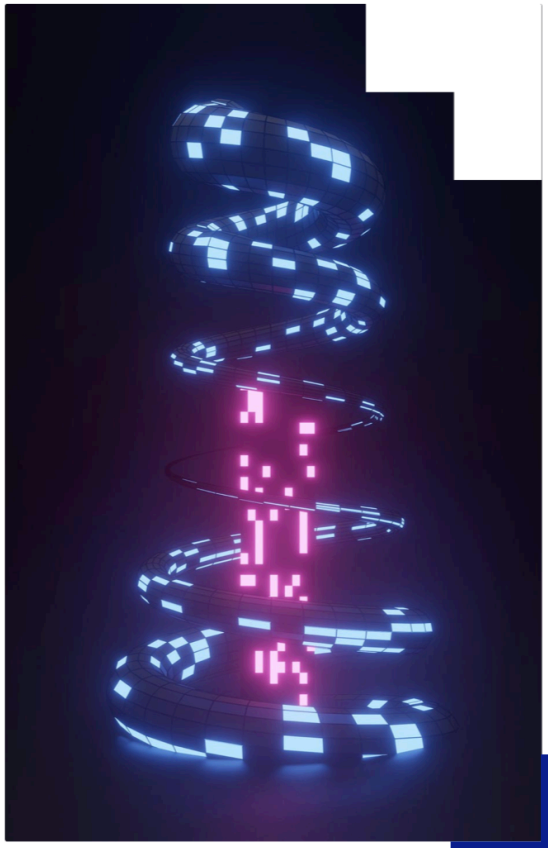
  Trained a baseline MLP and a cluster-based fairness-aware MLP model.
  Evaluated models using overall and fairness metrics across subgroups.

- **Fairness Stress Test Introduction**

  Conducted an out-of-distribution stress test by modifying the training dataset.
  Removed 1000 promoted female instances to simulate fairness robustness.

- **Re-evaluation of Models**

  Both models were re-evaluated under the new dataset to assess fairness.

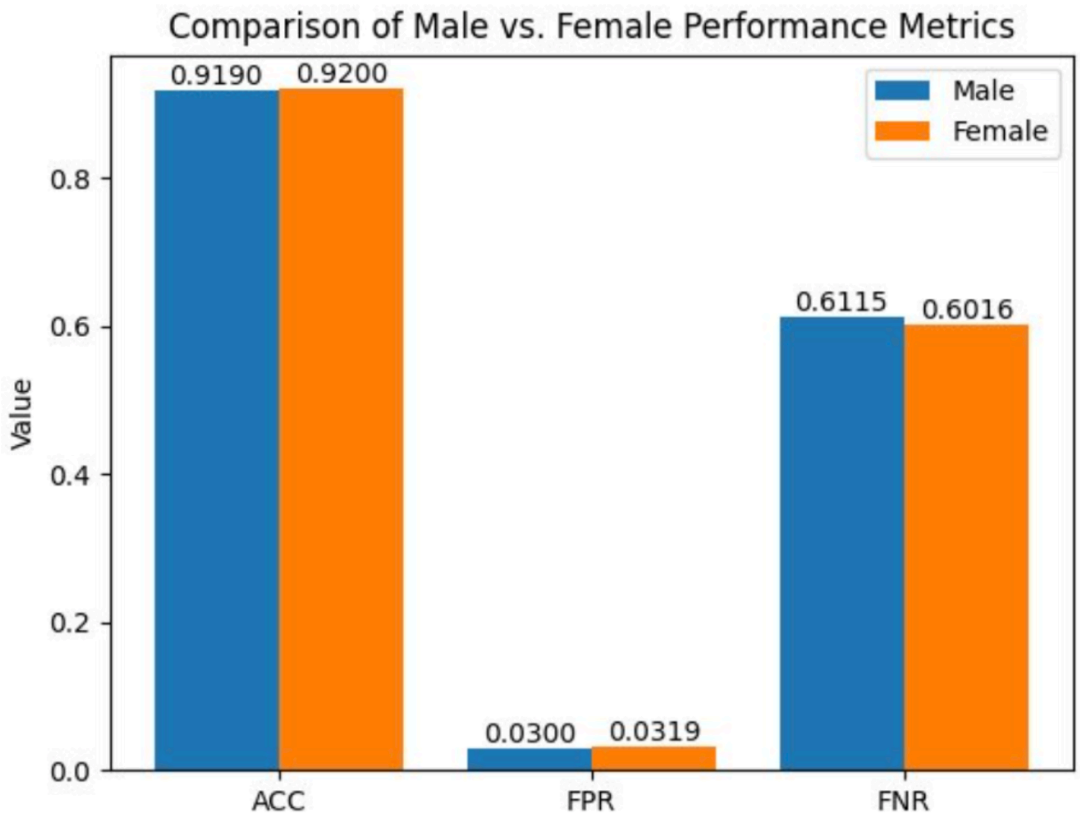


SECTION 03

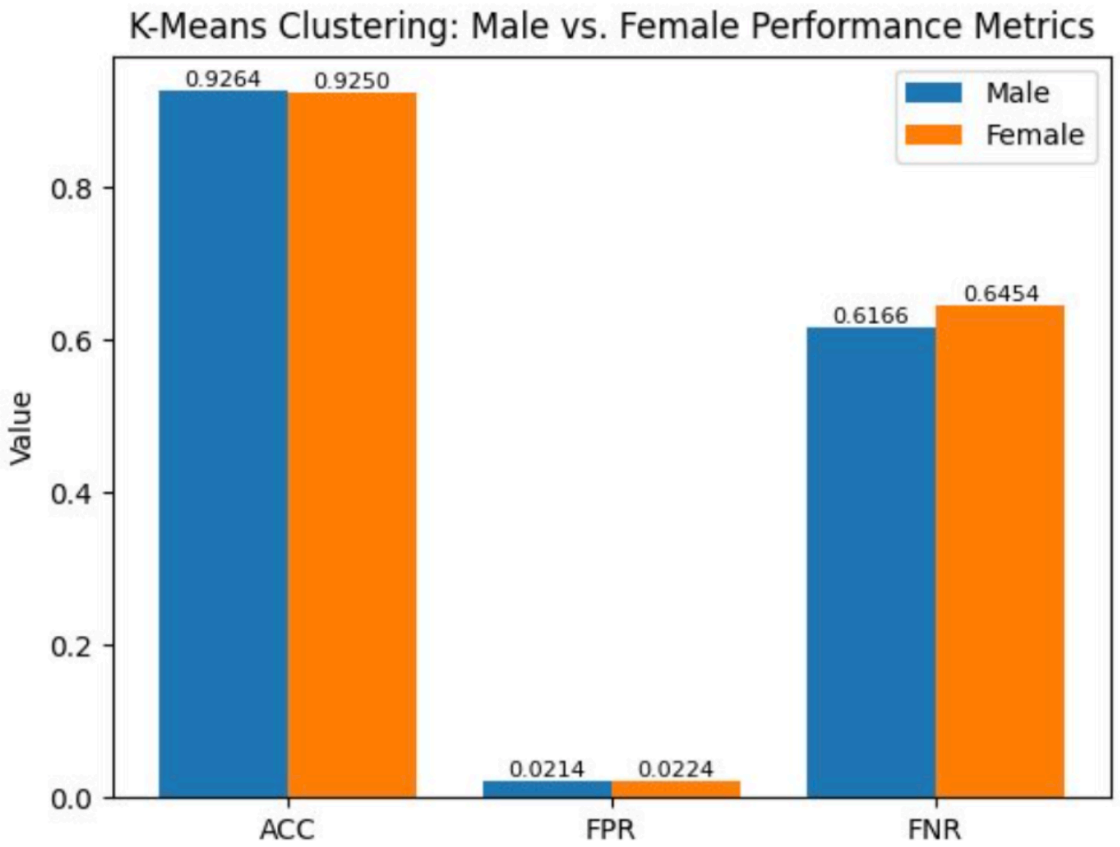
Results Overview

A comprehensive overview of the findings related to gender fairness in AI-powered employee promotion systems, highlighting key insights and data.

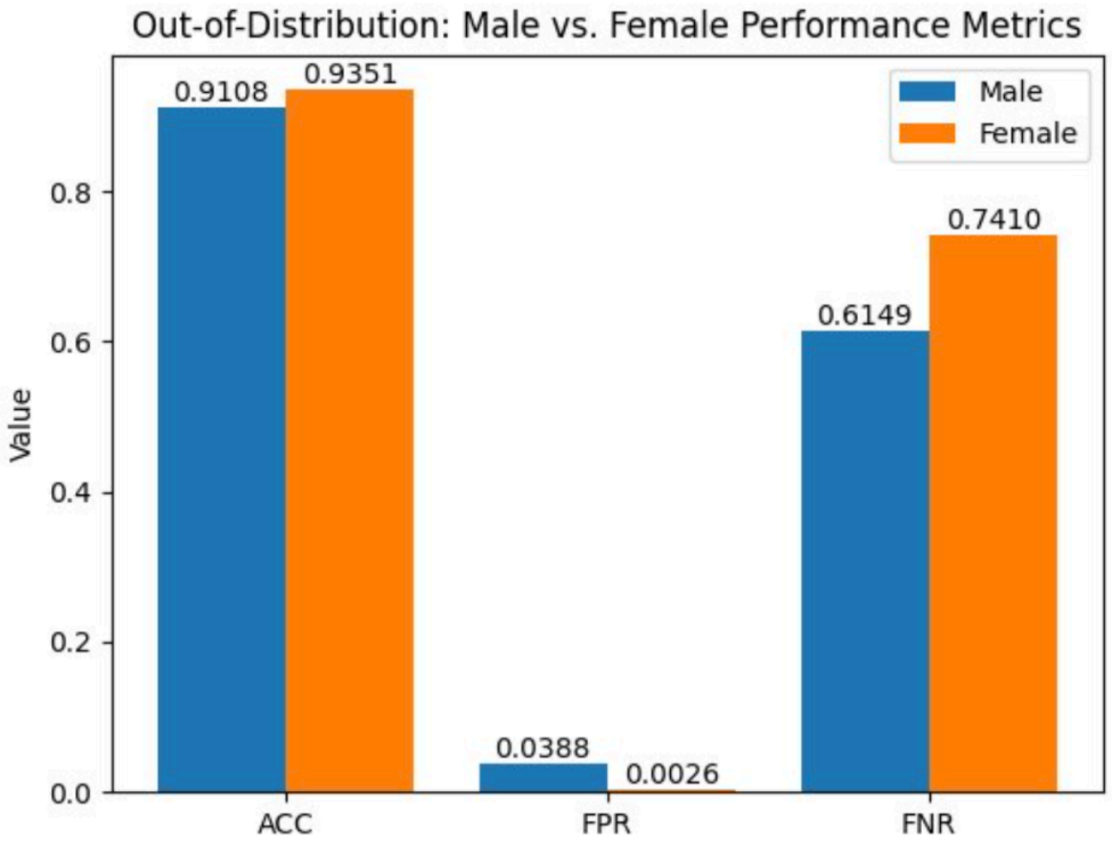
Subgroup Performance Comparison



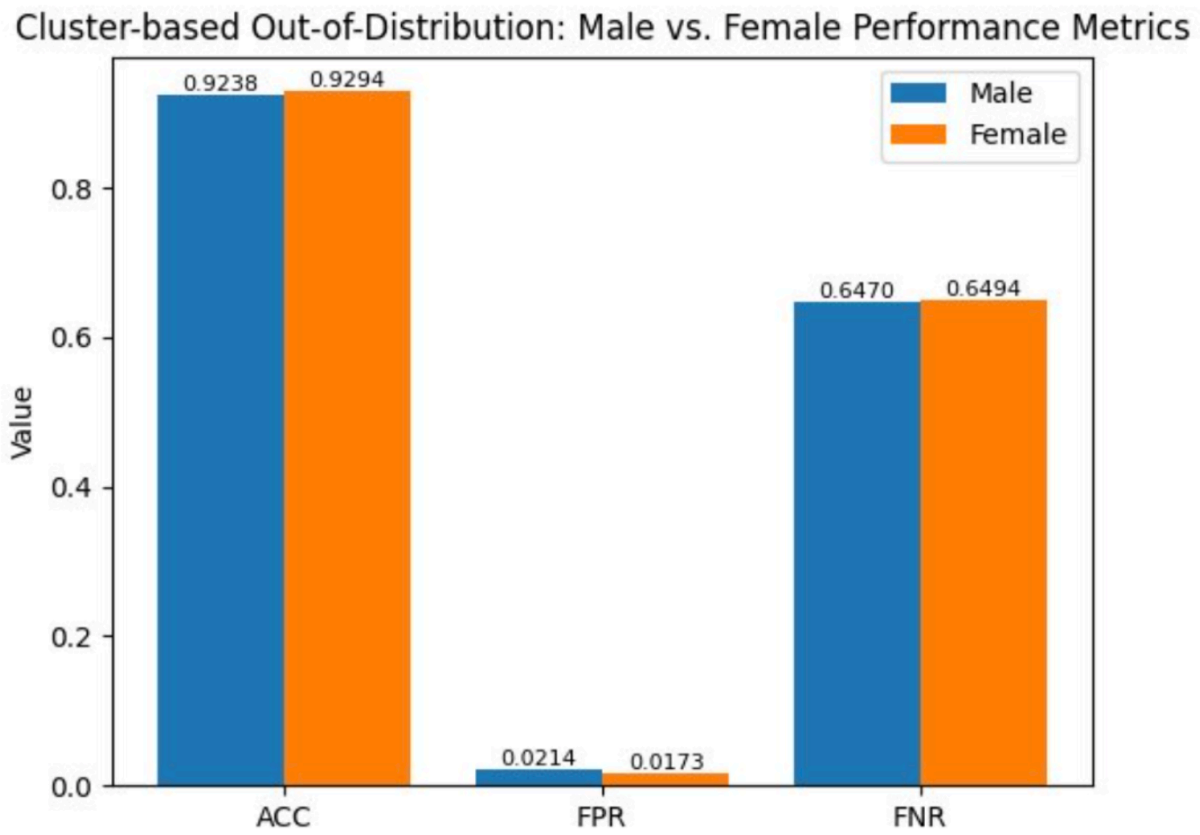
Subgroup Performance Comparison Using K-means



Subgroup Performance Comparison Using OOD



Subgroup Performance Comparison Using Cluster-based OOD



SECTION 04

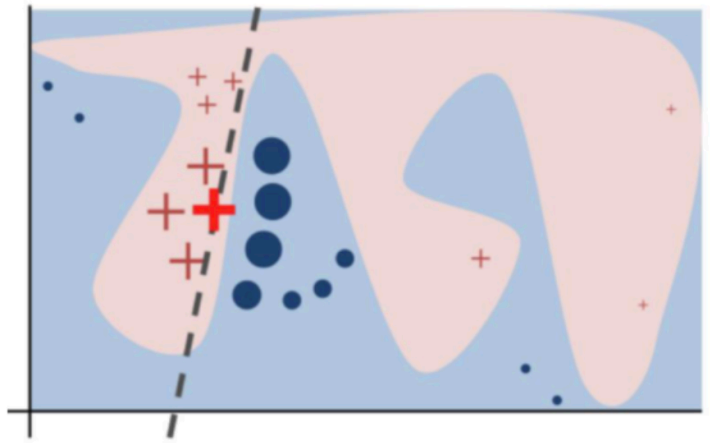
Further Experiment

What's more?

Further Experiment

Post-Processing – Explainability (XAI)

- Post-Processing step.
- Visualize Trustability of a Black-Boxed Model.
- Feature Importance.
- Training Data as a background of the data distribution
- Local Explanation with test data



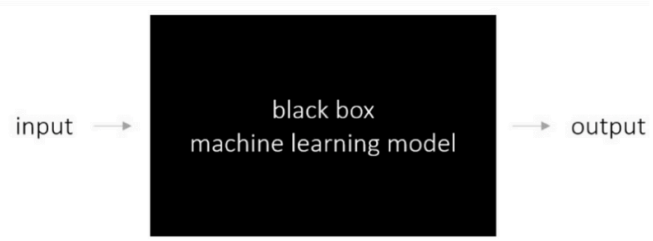
$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

- f(x): the complex model
- g(x): the simple & interpretable model (in our case, the MLP Classifier)
- Ω(g) : complexity of the simple model (can be still very complex)
- πx : Local region around x (data to be interpreted)

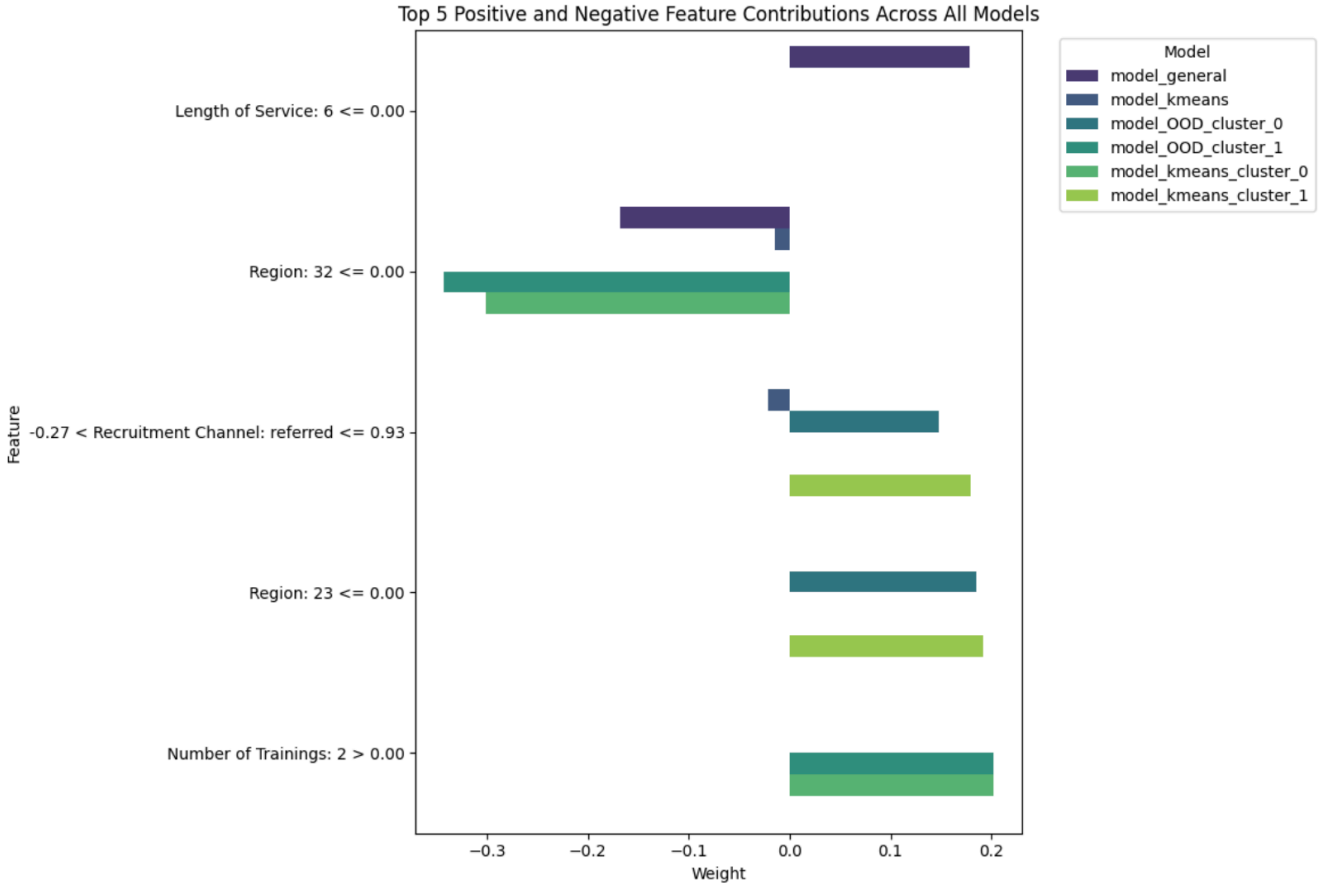
Further Experiment

XAI with LIME

- LimeTabularExplainer: provide context from training set.
- Explains the feature contributions in the test data.
- Relies on the model's predict_proba method.
- If the model doesn't have a predict_proba we customize it .
- Capable for any kinds of models (Scikit-Learn, TensorFlow, etc.)
- Scope of Explanation: 105 Features.



Feature Contribution Comparison Across Models



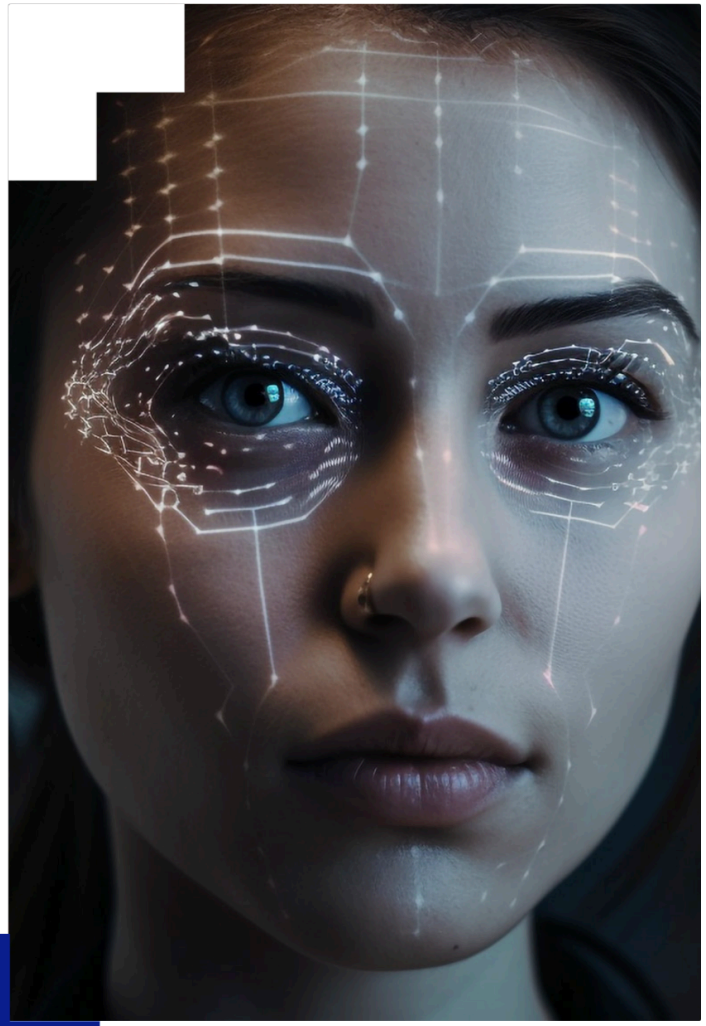
SECTION 05

Conclusion Overview

Highlights key learnings from our project results & how could we further extend our project?

Conclusion Overview

- **Fairness**: Equal Treatment Across Different Identifiable Groups!
- Our investigation revealed that AI Model could bring Fairness issue (First Model).
- New approach (second model) — including XAI tools such as **LIME** — we are able to:
  - Improve fairness across subgroups
  - Mitigate fairness problem in out-of-distribution situation
  - Provide transparency in decision-making
- Our project shows that **ethical AI in HR** is achievable but requires careful design, testing, and continuous improvement.
- Fairness is not just a technical goal—it's a **human responsibility**.



Future Work

Enhancing Fairness in AI Systems

- **Expand Dataset Scope**

  Utilize varied, industry-wide datasets to enhance fairness and diminish hidden biases.

- **Add Sensitive Attributes**

  Incorporate variables such as race, age, and disability status for comprehensive fairness assessments.

References

1. Dena F. Mujtaba and Nihar R. Mahapatra. “Ethical Considerations in Al-Based Recruitment” . Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8937920.

2. Dr. Changjian Shui. “CSI5195 Lecture 8 Explainability, Part 1 – Feature Importance” . University of Ottawa. 2025.

3. Gary K Y Chan. “Al employment decision-making: integrating the equal opportunity merit principle and explainable Al” . Al & SOCIETY Open Forum. 2024. Available at: https://link.springer.com/article/10.1007/s00146-022-01532-w.

4. Marco Tulio Correia Ribeiro. LIME PyPi Package. Available at: https://github.com/marcotcr/lime?tab=readme-ov-file.

FAIRNESS IN AI

Thank you

Know More

Github