Comparing Different Models' Reliance on Prohibited Features in the Adult Census Income

Daniele Baiocco
Ethics for AI

September 9, 2024

Abstract

The first goal of this project is to determine if a model's inherent feature importance serves as a proxy for its interpretability when assessed through Model-Agnostic Explanation methods. The second aim is to investigate whether low feature importance scores for prohibited features, obtained through model-agnostic explanation methods, correspond to a low Disparate Treatment Index [1]. If such a correspondence exists, it would suggest that models assigning low importance to these features are less likely to be discriminatory. Lastly, the project applies a Lagrangian approach to a neural network to minimize the Disparate Treatment Index, comparing feature importances before and after mitigation. This comparison highlights which features have decreased in importance, revealing those that were originally correlated with prohibited features.

1 Introduction

Machine learning models are increasingly used in high-stakes areas, raising critical questions about fairness and transparency. Understanding which features drive a model's predictions is essential for evaluating whether the model relies on prohibited features that could lead to discriminatory outcomes. While feature importance provides insights into which features influ-

ence predictions, it remains unclear whether low importance scores for sensitive attributes are enough to ensure a model's non-discriminatory nature. Moreover, even when prohibited features are excluded from the dataset, models can still exhibit biased outcomes due to the presence of other features that are correlated with these prohibited attributes. Identifying and understanding these correlated features is crucial for detecting and addressing hidden biases that may persist despite efforts to exclude sensitive information.

2 Background and Motivations

2.1 The Adult Income Dataset

The Adult Income Dataset [3], also known as the "Census Income" dataset, is widely used for machine learning and fairness analysis tasks. It contains demographic information on individuals, including attributes like age, education, occupation, race, sex, marital status, and work hours. The primary objective of the dataset is to predict whether an individual's annual income exceeds $50,000 based on these features.

2.2 Model Interpretability and Feature Importance

Different machine learning models offer various ways to assess feature importance. Below are the machine learning models I used:

- **Logistic Regression**: Provides explicit feature importance through the magnitude of its coefficients. Larger absolute values of coefficients indicate greater importance. Logistic Regression assumes a linear relationship between features and the target, making it straightforward to interpret the influence of each feature on predictions.
- **Gradient Boosted Decision Trees**: Models like XGBoost offer several approaches to measure feature importance:
  - **Weight**: Counts the number of times a feature is used to split nodes across all trees in the ensemble.
  - **Gain**: Measures the average gain in model performance attributable to a feature when it is used to split nodes.

  - **Cover**: Indicates the average number of samples affected by splits on a feature.
  - **Total Gain**: Summarizes the total gain in performance across all splits where the feature is used.
  - **Total Cover**: Summarizes the total coverage across all splits involving the feature.

- **Random Forests**: Feature importance is computed as the total reduction in the criterion (such as Gini impurity) brought by a feature, normalized across all trees. This measure, also known as Gini importance, quantifies how much each feature contributes to reducing impurity at the nodes where it is used to split. Features that lead to significant reductions in impurity are considered more important.

2.3 Model-Agnostic Techniques

In addition to model-specific feature importance metrics, model-agnostic methods like Permutation Importance [2] and LIME [4] (that are the ones used) offer valuable insights:

- **Permutation Importance**: Measures the impact of each feature on model performance by evaluating changes in accuracy when feature values are shuffled. A significant drop in performance indicates high importance.

- **LIME**: Provides local explanations by approximating the model with a simpler, interpretable model around a specific instance. This technique highlights which features most influence individual predictions.

2.4 Motivations

The motivations behind this project are the following:

- to investigate whether the inner model feature importance approximates the model-agnostic feature importance

- to investigate whether **low feature importance for prohibited attributes is sufficient to conclude that a model is non-discriminatory**

- to address the challenge of **identifying features indirectly correlated with prohibited attributes**. Understanding these indirect correlations is crucial to fully grasp which features contribute to discriminatory behavior, even when sensitive attributes are excluded from the dataset.

3 Project Description

3.1 Model Fitting and Evaluation

The project involves fitting four different models: Logistic Regression, Gradient Boosted Decision Tree, Random Forest, and a custom Neural Network. The models are trained on the Adult Census Income dataset, and their performance are evaluated using standard metrics to ensure that they accurately represent the underlying data. High evaluation metrics would confirm that the models were reliable estimators, suitable for further feature importance analysis.

3.2 Feature Importance Analysis

To interpret the models, the inner feature importances are extracted for Logistic Regression, Gradient Boosted Decision Tree, and Random Forest using their built-in interpretability methods. Permutation Importance and LIME are applied to all four models, with LIME global feature importance obtained by aggregating the feature importances across individual instances in the test set.

3.3 Experiment 1: Comparing Inner and Model-Agnostic Feature Importances

The first experiment aims to determine whether the inner feature importances of Logistic Regression, Gradient Boosted Decision Tree, and Random Forest align well with the model-agnostic feature importances obtained from LIME and Permutation Importance. For each of these models, a Cosine Similarity measure is computed between the model's inner feature importance arrays and the corresponding LIME and Permutation Importance arrays.

3.4 Experiment 2: Feature Importance of Prohibited Features

The second experiment focuses on evaluating the feature importance of prohibited attributes, specifically sex and race, using LIME and Permutation Importance for all four models. The experiment aims to identify which model exhibits the lowest feature importance scores for these sensitive features through visual analysis of feature importance plots. Subsequently, the Disparate Treatment Index (DIDI) is computed for each model, and comparisons are made to assess whether the model with the lowest feature importance for prohibited features also exhibits the lowest DIDI value. This analysis seeks to explore the relationship between low feature importance for prohibited features and reduced discriminatory behavior.

3.5 Experiment 3: Detect Feature Importances Correlated with Prohibited Features

In the third experiment, the focus shifts to the Neural Network model. A Lagrangian approach is employed to minimize the DIDI index, thereby aiming to reduce discriminatory outcomes. After training the mitigated Neural Network, Permutation Importance and LIME feature importances are recalculated. The Feature Importance arrays of the neural network before and after mitigation are then compared to identify features whose importance decreased. This comparison highlights features that are likely correlated with prohibited attributes, providing insights into which features contribute to biased model behavior.

4 Implementation

This section details the implementation of the project, including preprocessing steps, model training, and the application of bias mitigation techniques. The implementation was carried out using Python, primarily utilizing libraries such as scikit-learn, XGBoost, lime and Tensorflow.

4.1 Data Preprocessing

The Adult Census Income dataset underwent several preprocessing steps to prepare it for modeling:

1. Rows with missing values (denoted by ?) were removed to ensure data quality.

2. The ‘race‘ feature, originally containing multiple categories (‘White‘, ‘Black‘, ‘Asian-Pac-Islander‘, ‘Amer-Indian-Eskimo‘, ‘Other‘), was converted into a binary variable, with ‘1‘ assigned to ‘White‘ and ‘0‘ to all other categories.

3. The ‘marital status‘ feature was simplified into a binary variable: ‘0‘ was assigned to ‘Never-married‘, ‘Divorced‘, ‘Separated‘, and ‘Widowed‘, while ‘1‘ was assigned to ‘Married-civ-spouse‘, ‘Married-spouse-absent‘, and ‘Married-AF-spouse‘.

4. The ‘relationship‘, ‘native-country‘, ‘education-num‘, and ‘occupation‘ columns were dropped to reduce dimensionality.

5. The ‘workclass‘ and ‘education‘ features were encoded as ordinal variables to capture their inherent order.

6. Finally, all features were normalized to ensure that the models could train efficiently without being biased by feature scale differences.

4.2 Model Training and Hyperparameter Tuning

Three interpretable models—Logistic Regression, Random Forest, and Gradient Boosted Decision Tree—along with a custom Neural Network were fitted to the processed dataset. A hyperparameter tuning process was performed using Grid Search with cross-validation, optimizing for the ROC AUC score:

- **Logistic Regression:** Penalty type and regularization strength were tuned to select the optimal model.

- **Random Forest:** The number of estimators, maximum tree depth, and whether to use bootstrapping were tuned to enhance model performance.

- **Gradient Boosted Decision Tree:** The number of estimators and regularization parameter (λ) were adjusted to improve the model's prediction accuracy.

The custom Neural Network architecture consisted of a single hidden layer with 16 units, using the sigmoid activation function in the output layer. The model was trained with the Adam optimizer, using a learning rate of 0.0001.

4.3 Disparate Impact Discrimination Index (DIDI) Calculation

The Disparate Impact Discrimination Index (DIDI) was used to measure the discriminatory effect of the models. It was computed using the formula:

$$\text{DIDIc}(D) = \sum_{y \in Y} \sum_{x_p \in X_p} \left| \frac{|\{i \in N : y_i = y\}|}{|N|} - \frac{|\{i \in N : y_i = y \cap x_{p,i} = x_p\}|}{|\{i \in N : x_{p,i} = x_p\}|} \right|$$

Here, N represents the total number of instances, Y is the set of possible outcomes, and X_p is the set of prohibited attributes. The formula calculates the absolute difference between the overall probability of each outcome and the probability conditioned on each prohibited attribute. This measure highlights discrepancies in model behavior across sensitive groups.

4.4 Mitigation with Lagrangian Approach

To reduce the DIDI index and mitigate discriminatory behavior, a Lagrangian approach was applied to the Neural Network. The method involved optimizing a combined loss function:

$$\arg\min_{\theta} E[L(y, f(x, \theta)) + \lambda \max(0, \text{DIDI}(f(x, \theta)) - \epsilon)]$$

Here, λ is a learnable parameter that adjusts the penalty for DIDI violations, and ϵ is a tolerance threshold. The training process worked as follows:

- λ was initialized to 0.
- For each batch (equal to the entire dataset), the DIDI index and the binary cross-entropy loss were computed. The total loss was then used to update the model weights through a gradient descent step.

- After updating the model weights, the DIDI index was recalculated, and a gradient ascent step was performed on λ. This adjustment increased λ whenever a DIDI violation occurred, thus pushing the model to reduce DIDI in subsequent updates.

This iterative process allows the model to learn to minimize discriminatory behavior by dynamically adjusting its emphasis on fairness during training, leading to more equitable outcomes.

5 Results and Discussion

Table 1 presents the evaluation results of the four models—Logistic Regression, Gradient Boosted Decision Tree, Random Forest, and a Custom Neural Network—on the test split. All models exhibit high performance, confirming their reliability as estimators suitable for feature importance analysis.

Table 1: Performance Metrics of Different Models

Metric	**LogReg**	**RandFor**	**GradBoost**	**NNModel**
Accuracy	0.8410	0.8580	0.8619	0.8438
Precision Macro	0.7926	0.8324	0.8218	0.7969
Precision Weighted	0.8331	0.8524	0.8566	0.8362
Recall Macro	0.7463	0.7537	0.7827	0.7503
Recall Weighted	0.8410	0.8580	0.8619	0.8438
F1 Macro	0.7645	0.7813	0.7992	0.7688
F1 Weighted	0.8342	0.8486	0.8575	0.8372
AUC	0.8924	0.9134	0.9198	0.8951

5.1 Comparing Inner and Model-Agnostic Feature Importances

Figure 1 shows the cosine similarity between the inner feature importances and model-agnostic methods like *Permutation Importance* and *LIME*. The results indicate that, for most features, the inner feature importances closely match those from the model-agnostic approaches, suggesting that inner model feature importances can be a reliable proxy for these interpretability measures.

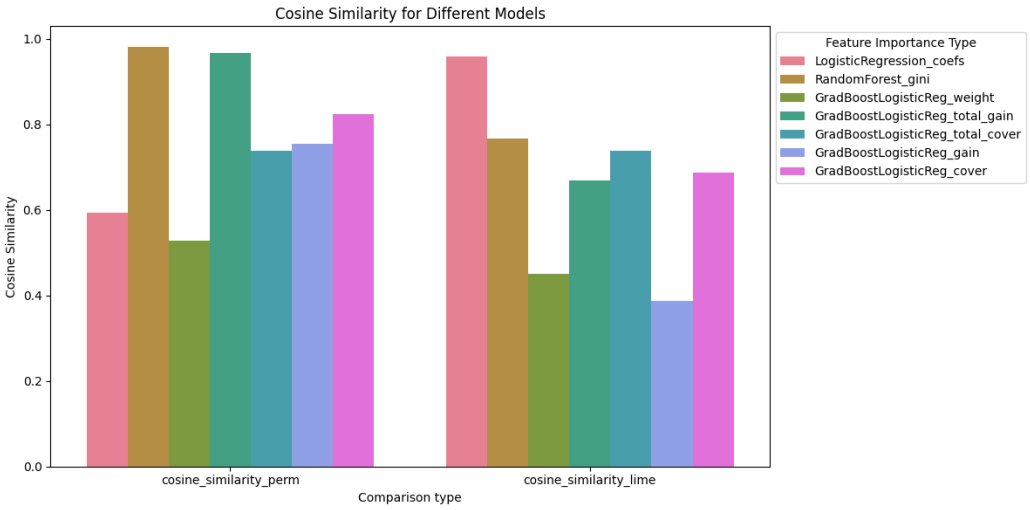


Figure 1: Cosine Similarity between different types of models feature importances and Model-agnostic feature importances computed with Permutation Importance and Lime

5.2 Feature Importance of Prohibited Features

Figure (2a) and Figure (2b) illustrate the feature importances of prohibited attributes (race and sex) as measured by LIME and Permutation Importance, respectively. In Figure (2a), all models show a zero importance for the race feature, while the sex feature importance is lowest for Random Forest and Gradient Boosted Decision Tree, moderate for Logistic Regression, and highest for the Neural Network model.

Figure 2b further shows that Logistic Regression and Random Forest have the lowest importance values for both race and sex, while Gradient Boost LogReg and Neural Network display higher values for these features.

Table 2 highlights that the Random Forest model consistently shows the lowest Disparate Treatment Index (DIDI) across both training and test sets, correlating with its low feature importances for the prohibited features in both LIME and Permutation Importance. However, the feature importance analysis alone is not sufficient to conclude a model's discriminatory nature; for example, Logistic Regression shows low feature importances but has a high DIDI score, suggesting potential biases that are not captured by feature importance metrics.

Model	DIDI_tr	DIDI_test	DIDI_tr_race	DIDI_test_race	DIDI_tr_sex	DIDI_test_sex
Dataset	0.610489	0.586485	0.212049	0.191358	0.398440	0.395128
LogReg	0.613727	0.571183	0.207201	0.184680	0.406526	0.386504
RandFor	0.430438	0.425889	0.126626	0.130848	0.303812	0.295041
GradBoost	0.576449	0.583375	0.206102	0.215640	0.370346	0.367735
NNModel	0.623481	0.580203	0.235734	0.206627	0.387747	0.373576

Table 2: DIDI Scores for Various Models

5.3 Detect Feature Importances Correlated with Prohibited Features

To reduce discrimination, a neural network model was retrained with a DIDI threshold of 0.3 on the training set. Figures (3a) and (3b) show the comparison of model-agnostic feature importances (LIME and Permutation Importance) before and after mitigation. Notably, features such as capital gain and education decreased significantly, highlighting their correlation with the prohibited features. This suggests that the dataset mirrors social biases, where capital gains and education levels are lower for individuals of certain races and genders.

6 Conclusion

The study found that inner model feature importances are generally good proxies for model-agnostic feature importances such as LIME and Permutation Importance. However, relying solely on feature importance to evaluate discriminatory behavior is insufficient, as demonstrated by the discrepancies between feature importance values and DIDI scores, particularly for Logistic Regression.

Applying a mitigation approach to the neural network model revealed which features were strongly correlated with prohibited features, suggesting potential areas for bias reduction. These insights highlight the need for careful feature selection and bias mitigation techniques to achieve fairer predictive models.

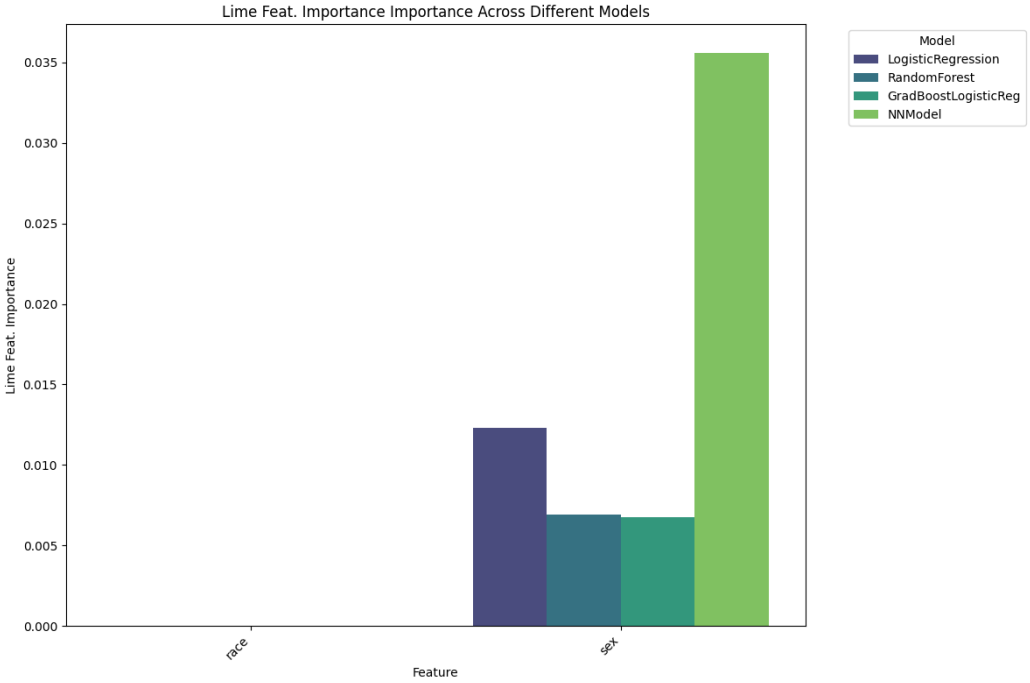
References

[1] Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos. Learning optimal and fair decision trees for non-discriminative decision-making. *Center for Artificial Intelligence in Society, University of Southern California*, 2024.

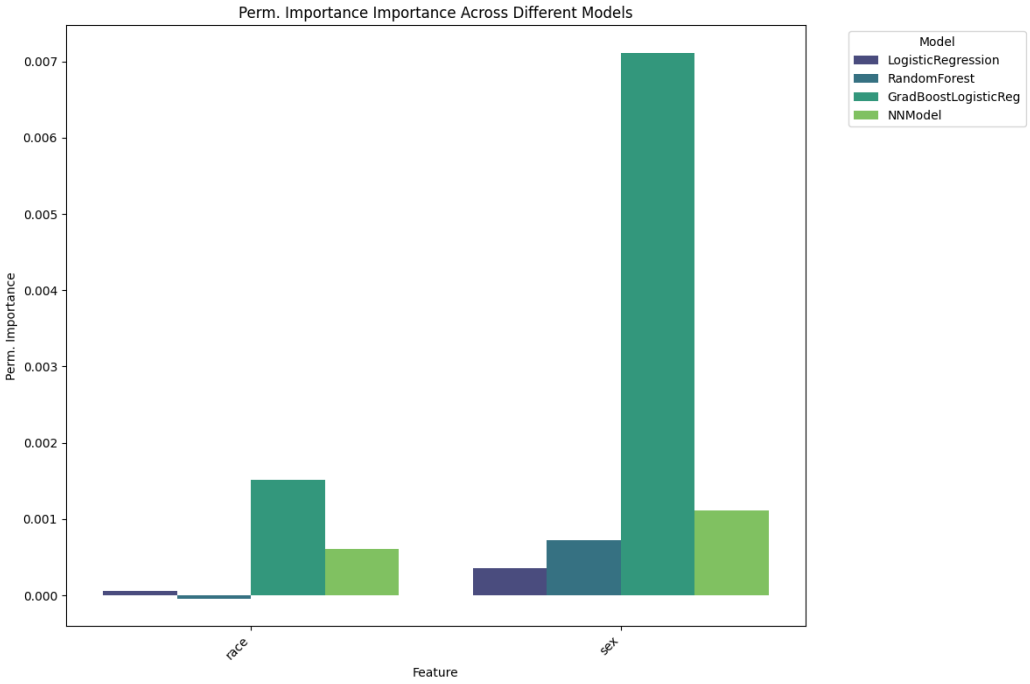
[2] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 04 2010.

[3] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. pages 1135–1144, 2016.

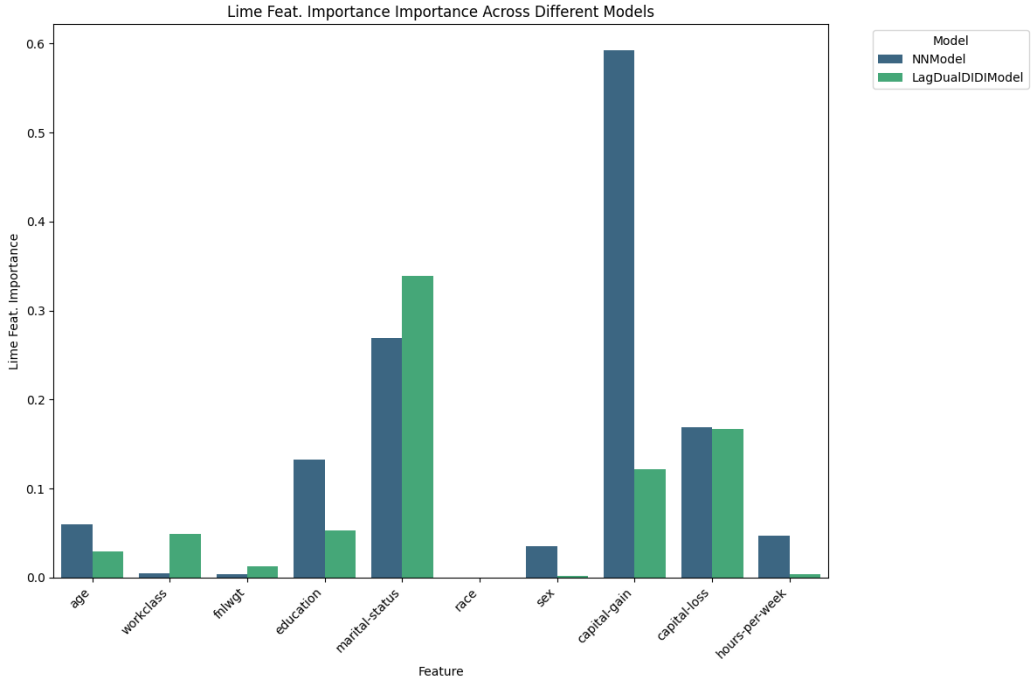


(a) LIME Feature Importances for prohibited features

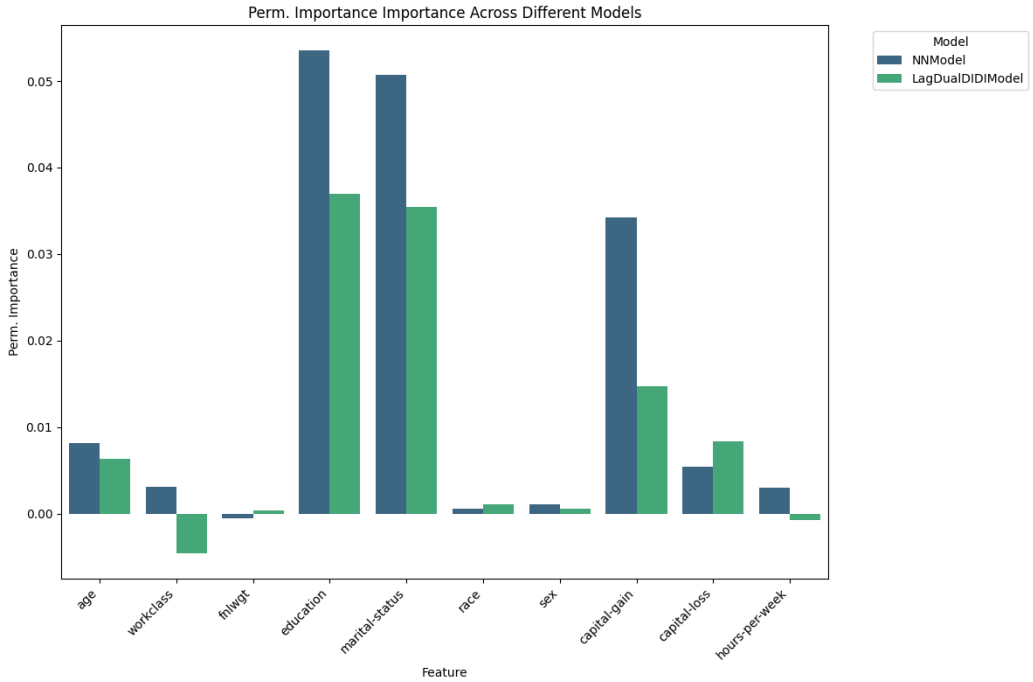


(b) Permutation Importances for prohibited features

Figure 2: The Figures contains respectively the LIME Feature Importances and the Permutation Importances of prohibited features race and sex across different models



(a) LIME Feature Importances before and after mitigation approach



(b) Permutation Importances before and after mitigation approach

Figure 3: The figures illustrate the LIME feature importances and Permutation Importances of the Neural Network model, both before and after applying the mitigation approach. These plots highlight which features are most associated with discriminatory outcomes, allowing us to identify the factors contributing to bias in the model's predictions.