Bias Mitigation in Automed Loan Eligibility Process

Ethics in Artificial Intelligence

Angileri, Marzi, Oshodi

Project description

- **Analyze a German credit risk dataset**

  Examine the dataset to understand patterns and trends in credit allocation.

- **Identify bias and discrimination**

  Detect any unfair practices or biases affecting certain groups in the credit allocation process.

- **Explore potential mitigation strategies**

  Investigate strategies to reduce or eliminate bias and discrimination.

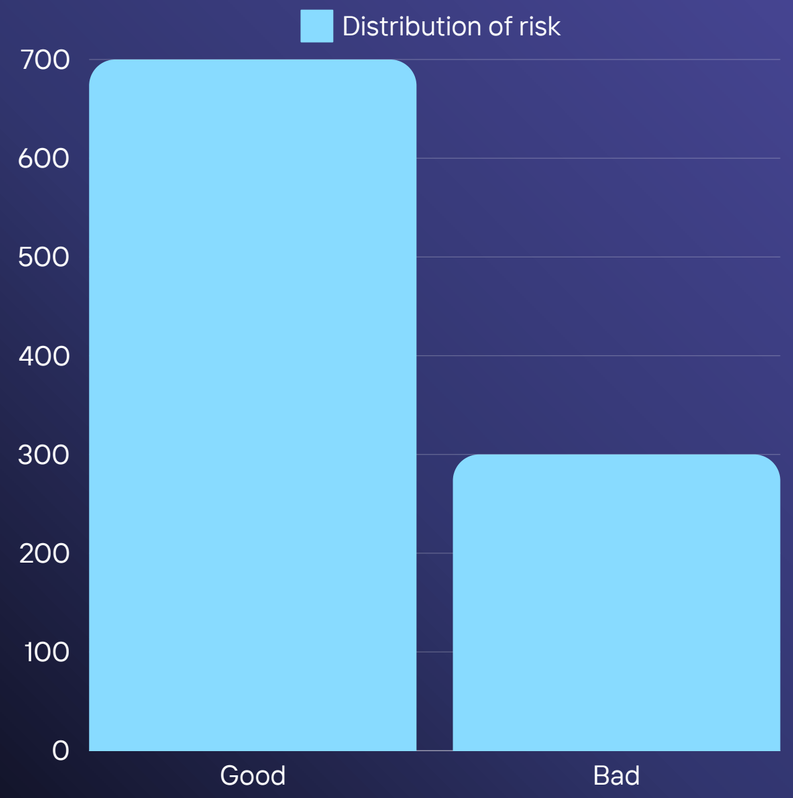
German Credit Risk Dataset

- 1000 samples
- 1 target variable
  - Risk (good - bad)
- 9 features
  - Age
  - Sex (male - female)
  - Job (0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled)
  - Housing (own, rent, or free)
  - Saving accounts (little, moderate, quite rich, rich)
  - Purpose (car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others)
  - Checking account
  - Credit amount
  - Duration (in month)

Univariate analysis

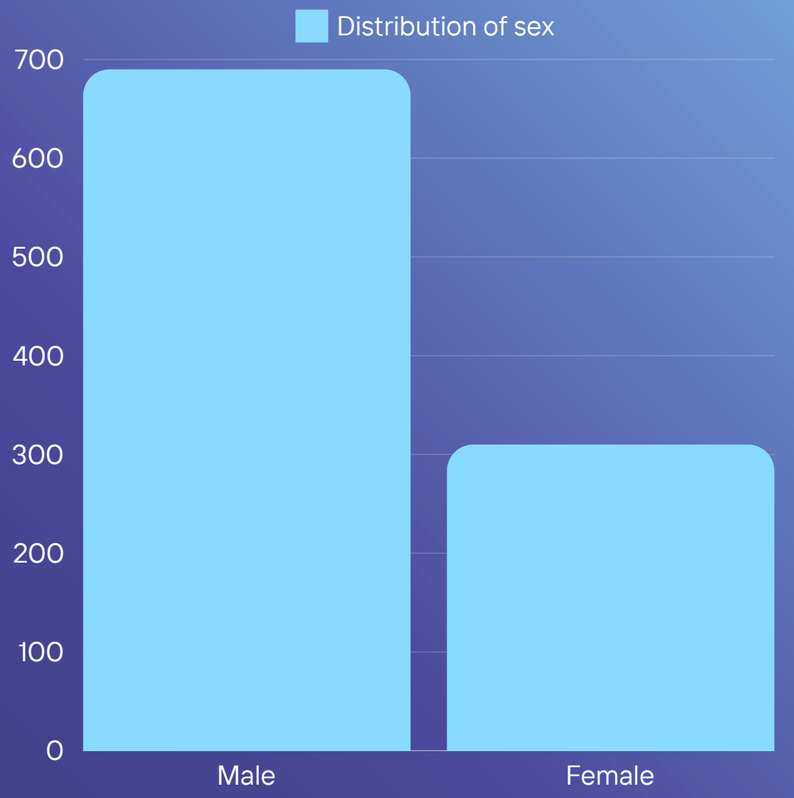
Target distribution

Instances of bad risk are relatively rare, making up less than half the number of good risk instances.



Sex distribution

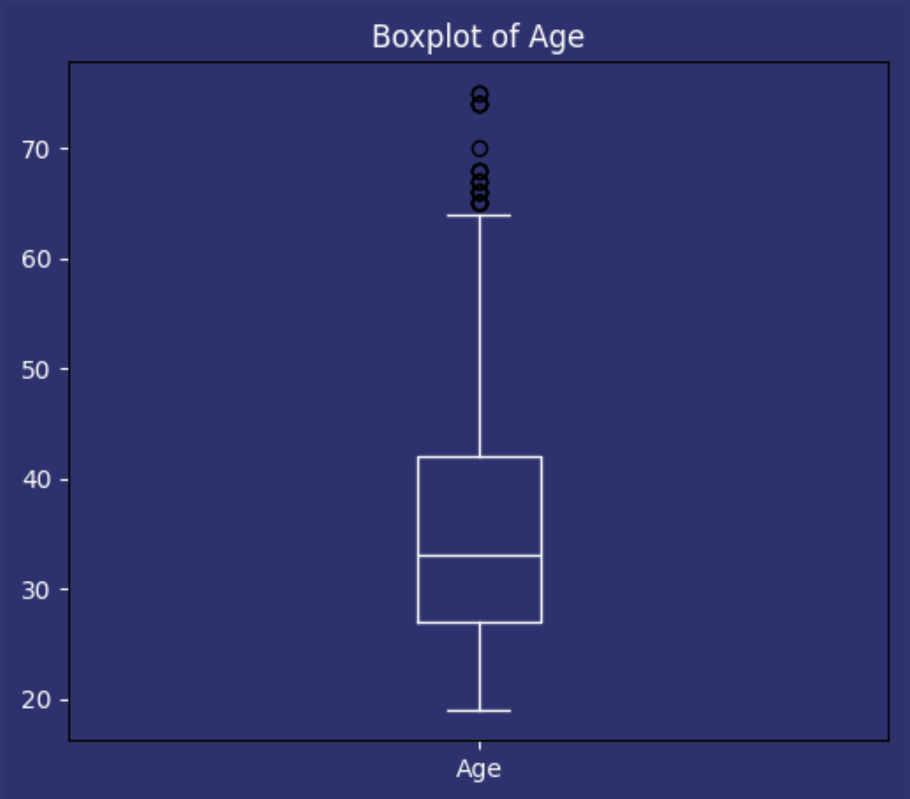
The dataset shows an imbalance in gender representation.



Univariate analysis

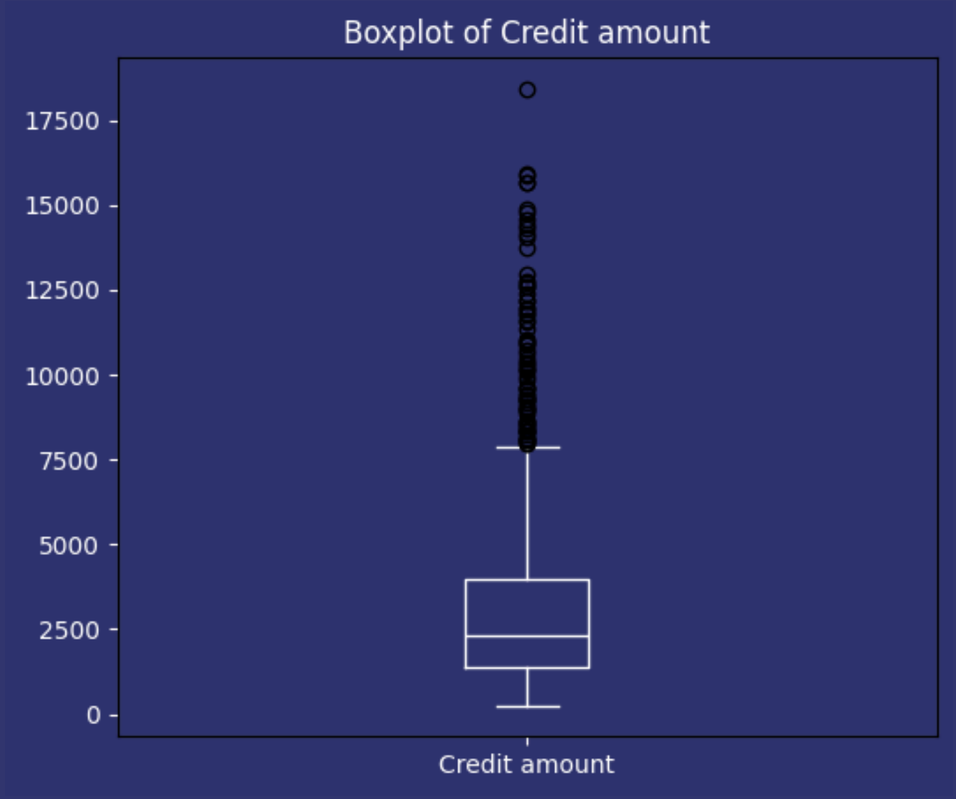
Age distribution

The age distribution is fairly normal with few outliers.

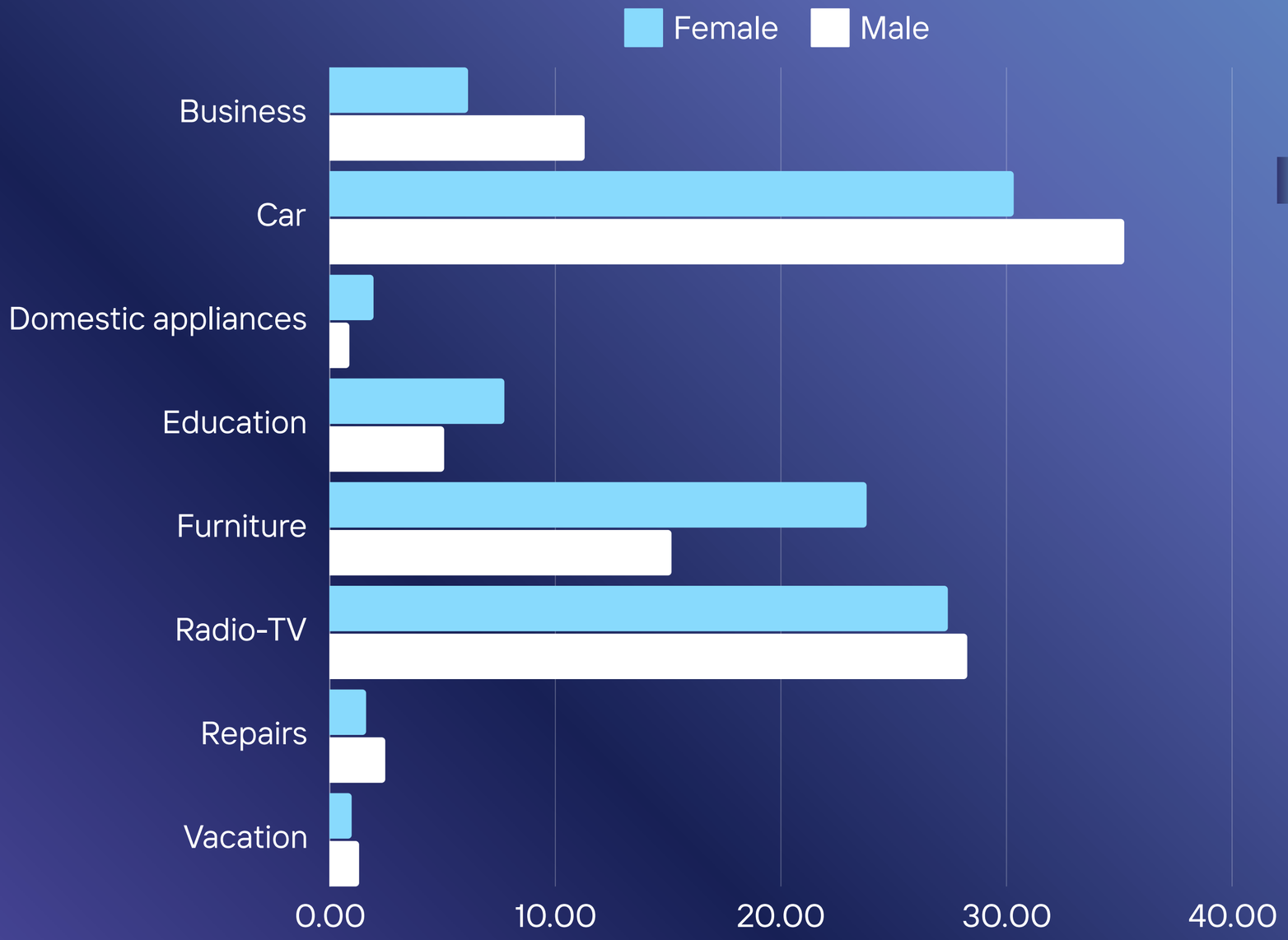


Credit amount distribution

The credit amount exhibits high variation and there are a lot of outliers.



Univariate analysis



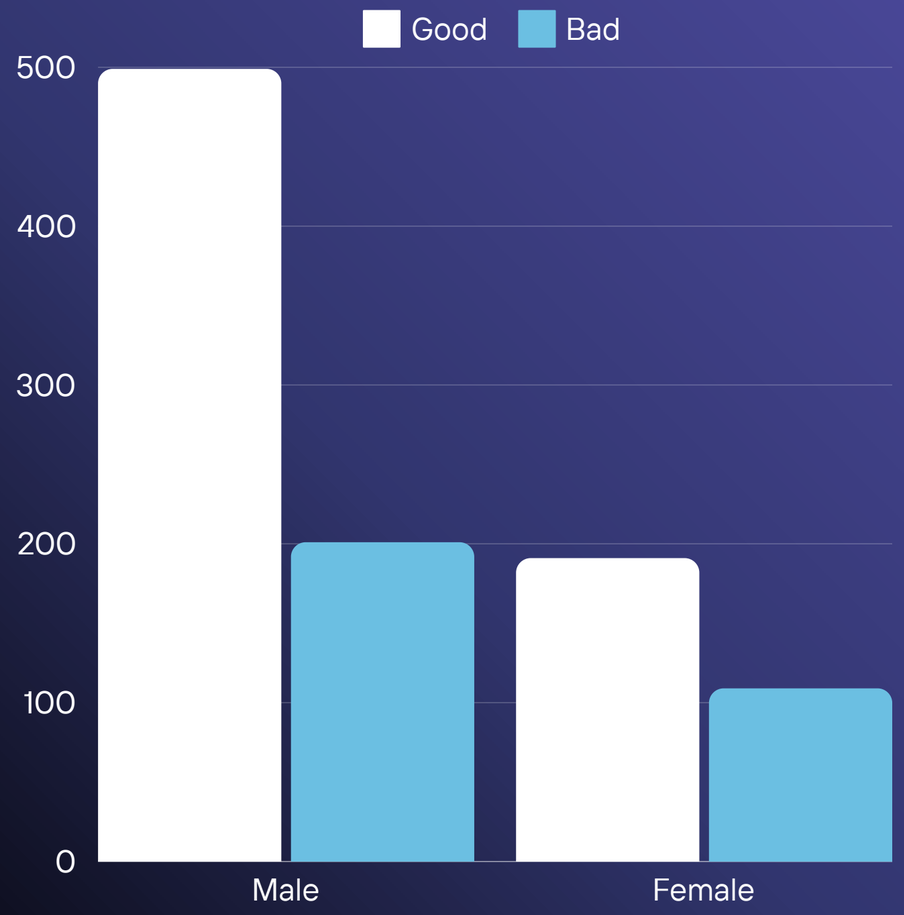
Purpose distribution

Further analysis aimed to uncover any differences in the purposes for which men and women applied for credit.

- Women skewed towards domestic appliances and forniture purchases.
- Men more prominent in business, repairs, and travel categories.

Bivariate analysis: Chi-square Test

The Chi-square Test determines association between categorical variables.



Risk vs Gender

- Significance level: 0.05
- p-value: 0.020
- The two variables Sex and Risk are dependent.

Risk vs Saving Accounts

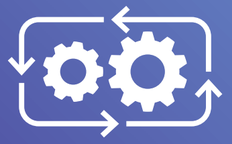
- Significance level: 0.05
- p-value: 0.0003
- The two variables Saving accounts and Risk are dependent.

Pipeline

Dataset Inspection and Preprocessing

Fairness metrics definition

Baseline Model



Mitigation Techniques

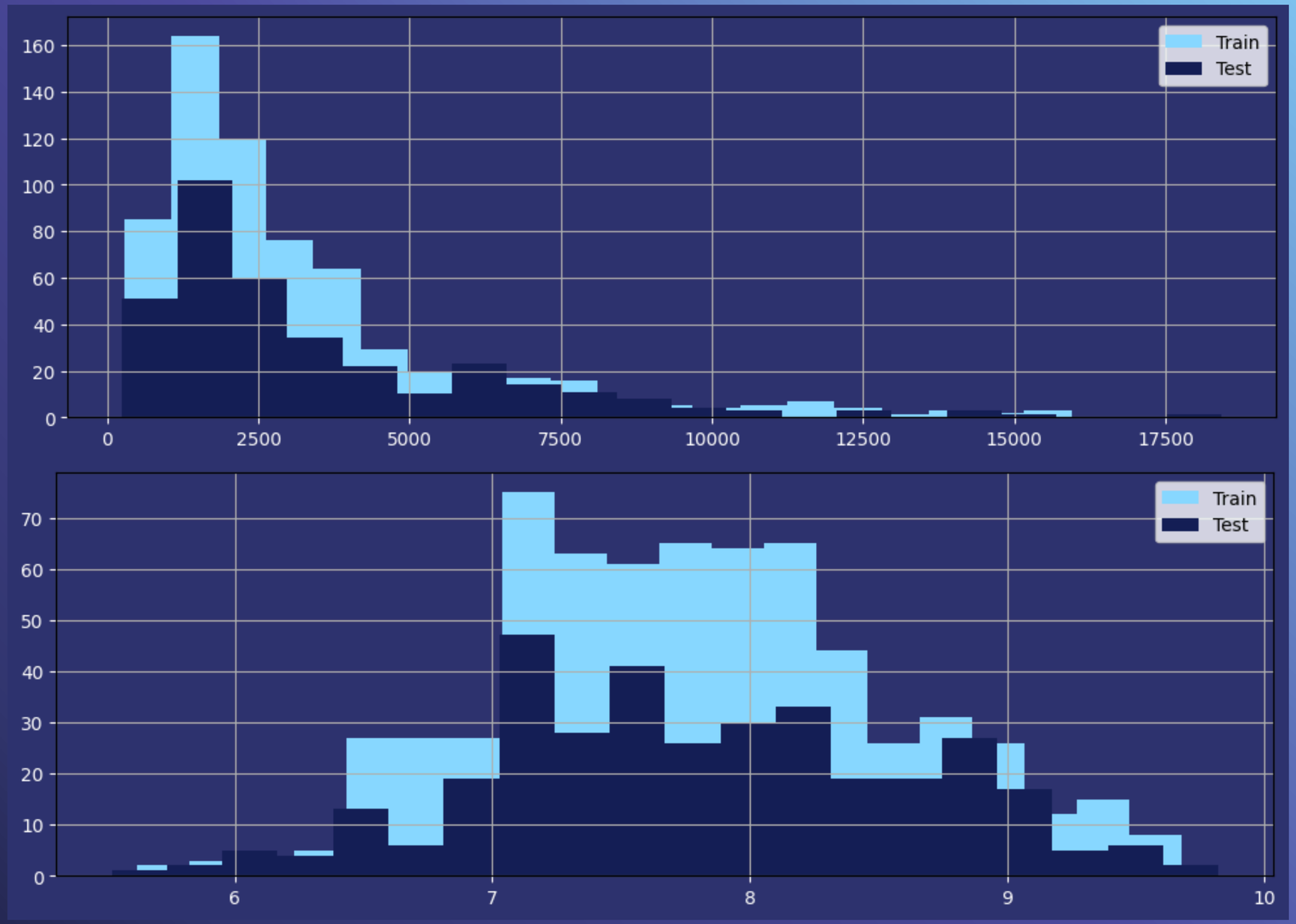
Comparison and Evaluation

Preprocessing

- **Age categorization**
  (Young, Adults, Senior, Elder)
- **Dropping columns**
  (Unamed:0, Age)
- **One Hot Encoding**
  (Job, Housing, Saving account, Checking account, Purpose, Age Group)
- **Outliers Removal**
  (Credit Amount)

Outlier Removal

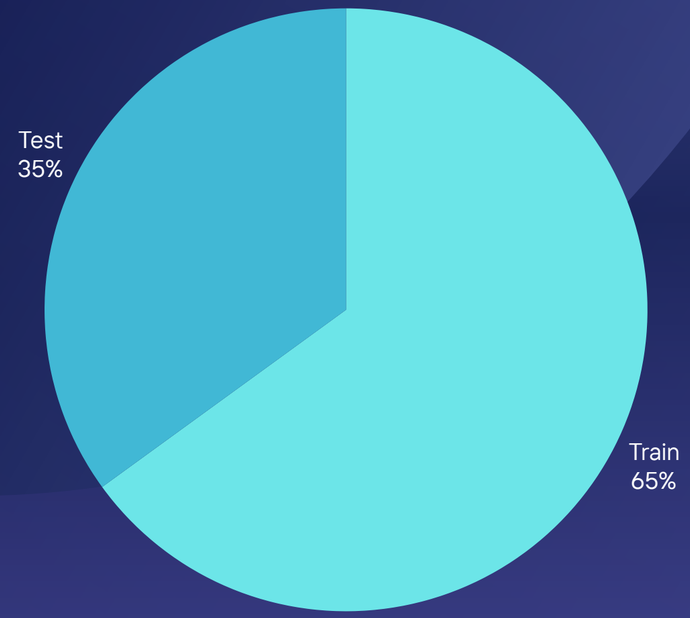
Credit Amount



Train-Test Split

- **Stratified sampling**
  - Same proportion of target

- **Sex: sensitive attribute**
  - Dropped from dataset
  - Used to evaluate intrinsic bias



Fairness Metrics

- **Demographic Parity**

  Each group should have an equal proportion of positive and negative predictions.

- **Equalized Odds**

  Each group should have an equal true positive rate and false positive rate.

- **Selection Rate**

  Positive predictions rate.

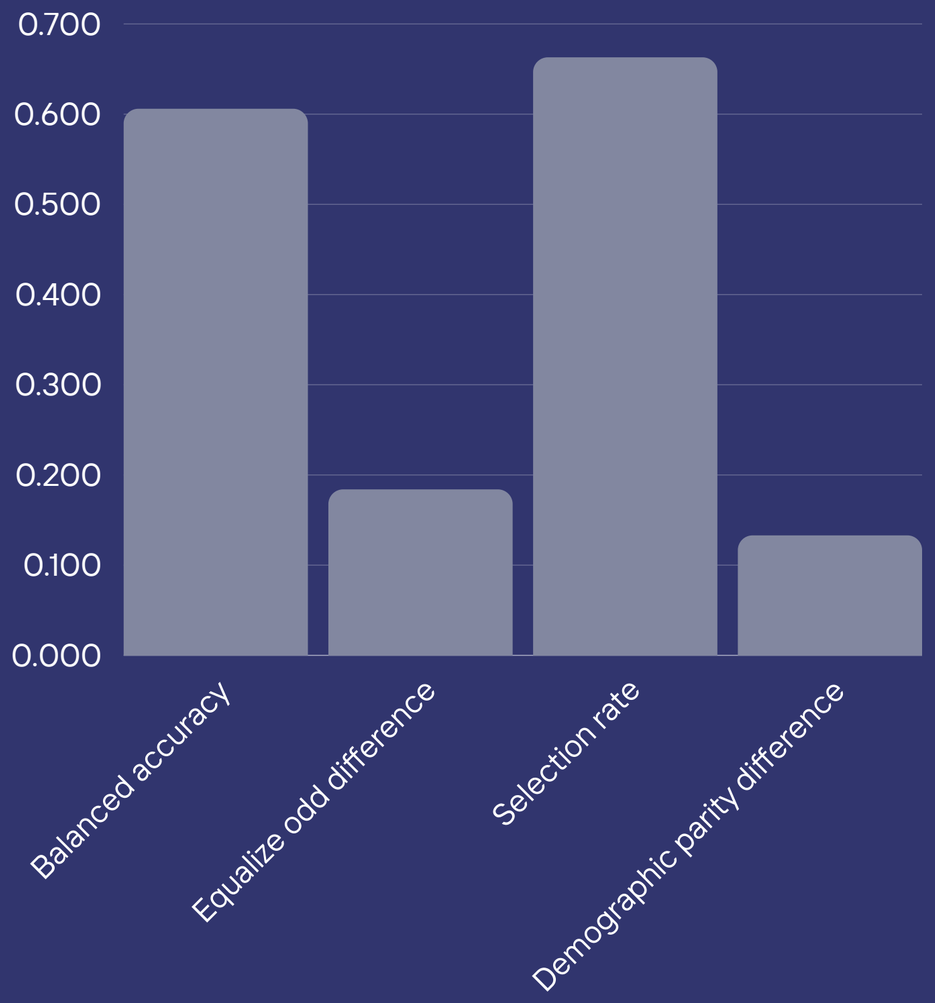
Statistical independence does not guarantee fairness

Similar people should have the same outcome

Baseline

Decision Tree Classifier

- Low accuracy but better than a random classifier
- Gender biases are present



Mitigation Techniques

Apply some in-process and post-process bias mitigation techniques and evaluate which approach better achieves our goals.

In-process

Exponentiated Gradient

Grid Search

Adversial Debiasing

Prejudice Removal

Post-process

Threshold optimizer

Exponentiated Gradient

InProcess mitigation - Iterative and exponential gradient-based approach to find an ensemble of models that optimize accuracy under fairness constraints (equalized odds in our case).

1. Uniform initialization of weights.
2. Iterations:
    a. Calculate the gradient of the loss with respect to the current weights, considering both the **loss** and the **fairness** metric.
    b. Updates the weights using an **exponential** factor based on the calculated gradient.
    c. Use the updated weights to train a **new** model.
    d. Combine models trained in previous iterations into a weighted **ensemble**, where the weights are determined by exponential updating.
3. Until it reaches convergence.

Grid search

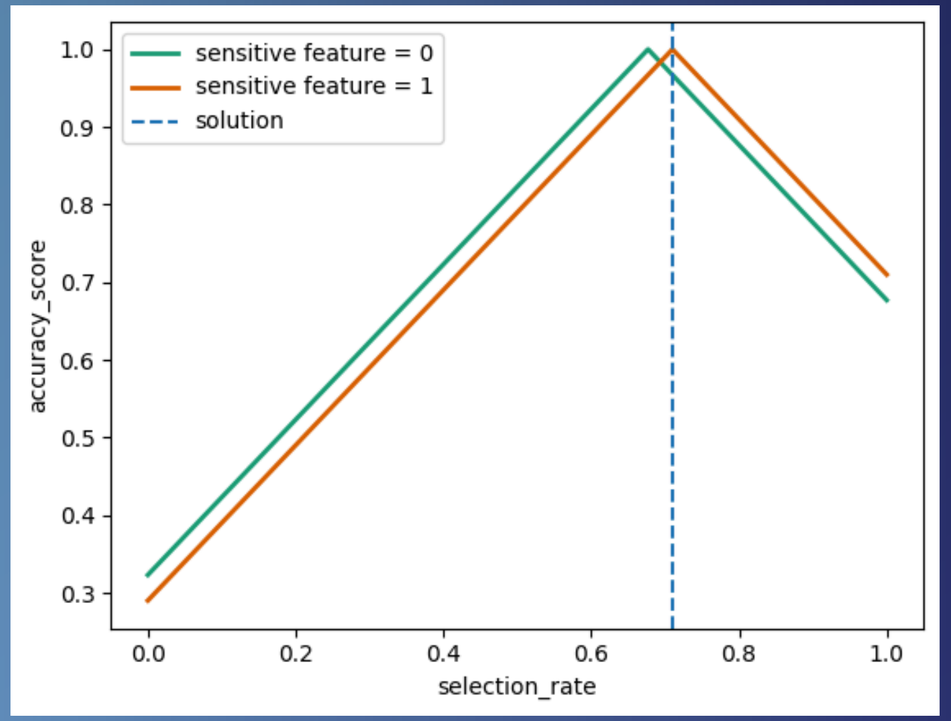
InProcess mitigation - Explore different model configurations and select the one that strikes the best balance between predictive performance and fairness metrics.

- Broader exploration of the parameters space.
- 600 different configurations explored.
- The quality of the solution depends on the granularity of the grid. A grid that is too coarse may not find the optimal solution.

Threshold Optimizer

PostProcess mitigation - Adjusts the **decision threshold** of model's predicted probabilities for fairness criteria.

- **Predicted Probabilities**: Model generates probabilities for each class.
- **Threshold Optimization**: Algorithm adjusts classification threshold based on fairness metrics.
- **Evaluation and Refinement**: Model's performance evaluated using fairness and traditional metrics.
- **Iterative Process**: Involves adjusting thresholds, evaluating performance iteratively until desired outcomes.

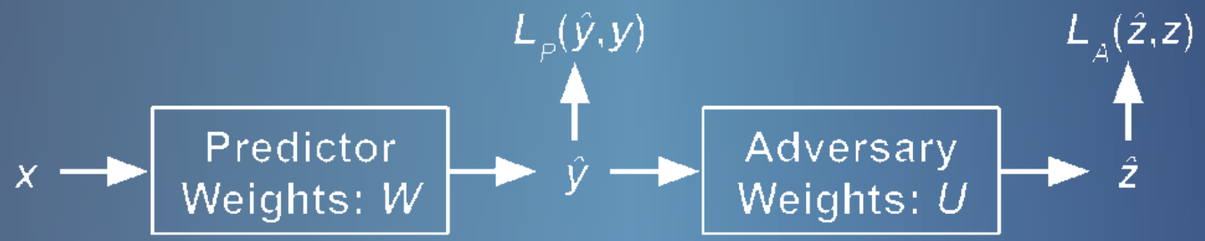


Prejudice Removal

InProcess mitigation - Reduces the statistical dependence between sensitive features and the remaining information by the addition of a fairness term to the **regularization parameter** that avoids over-fitting.

- **Regularization term** quantifies and penalizes decisions contributing to discrimination or bias, taking into account the senitive attribute.
- Encourages the model to learn **fair decision boundaries across demographic groups.**
- **Integrated into the training process** to promote fairness and equity in predictions.

Adversial Debiasing



- The **predictor** is trained to minimize its loss improving prediction accuracy.
- The **adversary** is trained to maximize its loss enhancing its ability to detect bias.
- Additionally, the **predictor** adjusts its weights to minimize the **adversary's** ability to correctly infer the protected attribute, thereby **reducing bias**

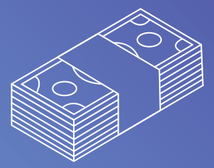
Risk Prediction Evaluation

	Unmitigated	GridSearch	Threshold Optimizer	Exponentiated Gradient	Prejudice Remover	Adversial Debiasing
Precision Male	0.825	0.817	0.825	0.825	0.814	0.806
Precision Female	0.621	0.648	0.640	0.621	0.671	0.644
Recall Male	0.778	0.784	0.778	0.778	0.920	0.948
Recall Female	0.594	0.696	0.695	0.594	0.739	0.840

Fairness Comparison

	Unmitigated	GridSearch	Threshold Optimizer	Exponentiated Gradient	Prejudice Remover	Adversial Debiasing
Balanced Accuracy	0.606	0.608	0.610	0.606	0.639	0.616
Equalize Odds Difference	0.184	0.088	0.095	0.184	0.181	0.108
Selection Rate	0.662	0.694	0.688	0.662	0.785	0.848
Demographic Parity Difference	0.132	0.132	0.132	0.132	0.132	0.132

Final Analysis



GridSearch and Threshold Optimizer emerge as the most effective techniques in this scenario, likely due to GridSearch's ability to explore a wider range of configurations and Threshold Optimizer's effectiveness in selecting the best decision threshold.

Adversial Debiasing performance falls between GridSearch and Threshold Optimizer, suggesting it might be less exhaustive in its search but still effective.

Prejudice Remover and Exponentiated Gradient show minimal bias reduction, likely due to less exhaustive parameter exploration or sensitivity issues.

Thank You!