

# Example Fairness report PDF




Fork FairnesLens here: <https://github.com/Bence749/FairnessLens>

Requirement already satisfied: pip in /Users/bencezsoltberegi/miniconda3/envs/Final\_P  
ER/lib/python3.12/site-packages (25.0.1)

Grid search training completed in 5.08 seconds

## FairnessLens: A Guide for Users on Detecting Bias and Fairness in Datasets and Models

Welcome to FairnessLens! This guide will walk you through the process of using our tool to detect potential bias and fairness issues in your dataset and model. This tool focuses on analyzing your dataset in both pre- and post-processing stages of the machine learning lifecycle.

Fairness Metrics Summary			
 Metric	Value 		 Description
Correlation average	0.0884		Measures the correlation between protected attribute and model predictions
Statistical Parity average	0.1125		Ensures equal probability of positive outcome across groups
Equalized Odds	age	personal_status_sex	Ensures equal true positive and false positive rates across groups
	0.0000	0.5333	
Equal Opportunity	age	personal_status_sex	Ensures equal true positive rates across groups for qualified individuals
	1.0000	0.1765	

## Numerical Values Analysis

This table summarizes the numerical attributes in the dataset. It includes minimum, maximum, average, standard deviation, and median for each column.

What to look for & why:

- Large variations between min and max → Extreme differences may indicate outliers, which can skew model predictions and affect fairness. If certain groups have significantly different values, the model may disproportionately favor one over another.
- A high standard deviation → A wide spread of values suggests data inconsistency, which can make predictions less reliable, especially for underrepresented groups.


- A median far from the average → A strong difference between these values suggests skewed data, meaning the dataset is not evenly distributed. This can lead to biased models that perform well on one subgroup but poorly on others.

### Categorical Values Analysis




This table displays categorical attributes, highlighting the most common and least common values along with their counts.

What to look for & why:

- Highly imbalanced categories → If one category appears far more often than others, the model may become biased toward that dominant category, leading to unfair predictions for less frequent groups.
- Rare categories with very few instances → Underrepresented groups may not have enough data for the model to learn from, resulting in poor generalization and higher error rates for those groups.



Numerical Values Analysis

feature	Numerical Attributes				
	Mean	Std Dev	Median 	Max 	Min 
duration	20.90	12.06	18	72	4
amount	3,271.26	2,822.74	2,320	18,424	250
installment_rate	2.97	1.12	3	4	1
age	35.55	11.38	33	75	19
number_credits	1.41	0.58	1	4	1
people_liable	1.16	0.36	1	2	1
present_residence	2.84	1.10	3	4	1



## Categorical Values Analysis

feature	Categorical Attributes			
	★ Most Common	Count	⚠ Least Common	Count
status	no checking account	394	... >= 200 DM / salary for at least 1 year	63
credit_history	existing credits paid back duly till now	530	no credits taken/all credits paid back duly	40
purpose	domestic appliances	280	business	9
savings	... < 100 DM	603	... >= 1000 DM	48
employment_duration	1 <= ... < 4 years	339	unemployed	62
personal_status_sex	male : single	548	male : divorced/separated	50
other_debtors	none	907	co-applicant	41
property	car or other	332	unknown/no property	154
other_installment_plans	none	814	stores	47
housing	own	713	for free	108
job	skilled employee/official	630	unemployed/unskilled - non-resident	22
telephone	no	596	yes	404
foreign_worker	yes	963	no	37

## Attribute Distribution Plots

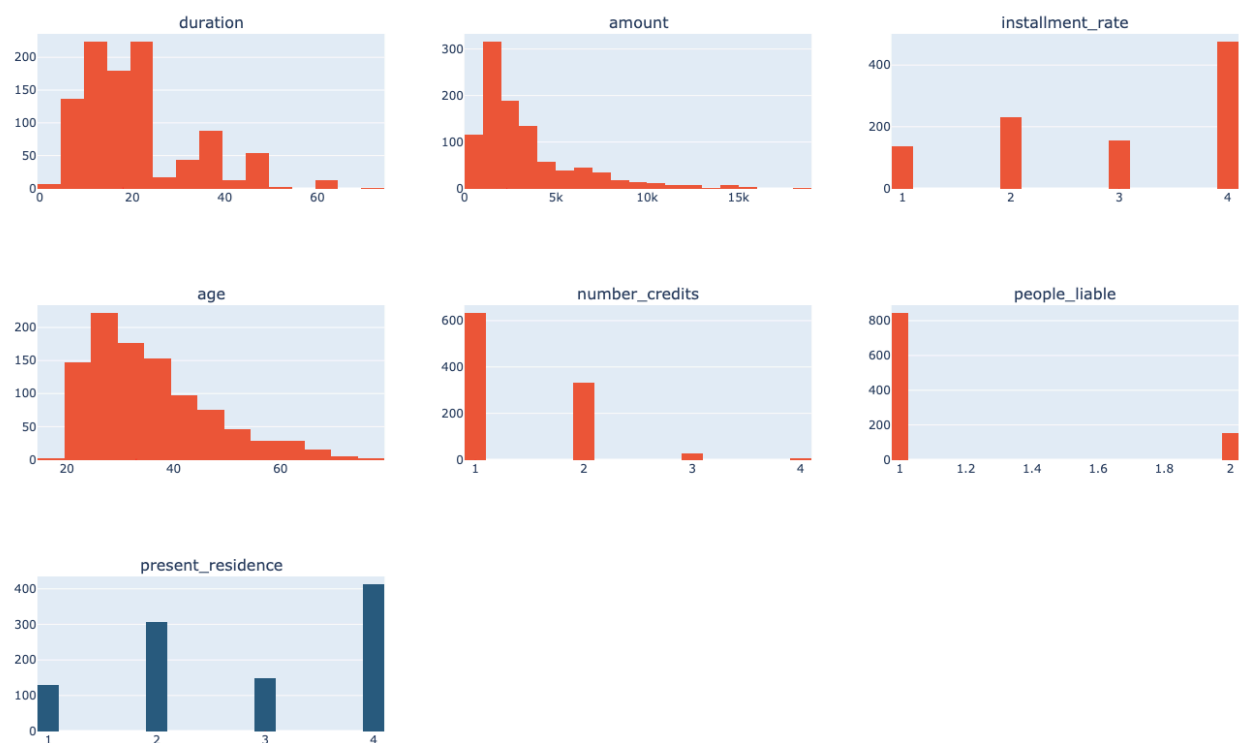
The following bar plots represent the distribution of each dataset attribute. Hovering over the bars reveals additional details, such as count and attribute names.

- Blue bars indicate a balanced distribution.
- Red bars highlight significantly unbalanced attributes, which may affect fairness.

What to look for & why:

- Attributes with red bars → A strongly skewed distribution means the dataset may not represent all subgroups fairly. If a feature is used in prediction and is highly imbalanced, the model may favor majority groups.
- Skewed distributions in sensitive attributes → If sensitive attributes like gender, race, or age show heavy imbalance, predictions may be unfairly biased. Addressing this through reweighting, balancing or additional data collection is recommended before training.

Numerical Feature Analysis



Categorical Feature Analysis



## Model Performance Analysis

After training a Grid Search model, additional graphs are generated to evaluate prediction patterns, class performance, and feature correlation. These insights help assess fairness and model behavior.

1. Prediction Patterns (Confusion Matrix) This matrix visualizes actual vs. predicted outcomes, where:
  - Rows represent actual values (true/false)
  - Columns represent predicted values (true/false)

- Darker blue indicates higher values (closer to 100), meaning stronger concentration in that category

#### What to look for & why:

- **More concentration on the diagonal (true positives & true negatives)** → Indicates good model performance, meaning correct predictions are frequent.
- **Off-diagonal values (false positives & false negatives) are high** → Suggests the model is making mistakes more often, which may lead to fairness issues.
- **More false positives (predicted true but actually false)** → If sensitive groups are affected, this could result in wrongful decisions (e.g., an unfair rejection or approval).
- **More false negatives (predicted false but actually true)** → Indicates missed opportunities, which may disproportionately impact underrepresented groups.

💡 Takeaway: If false positives or false negatives are significantly imbalanced across different groups, fairness adjustments may be required, such as reweighting data or adjusting decision thresholds.

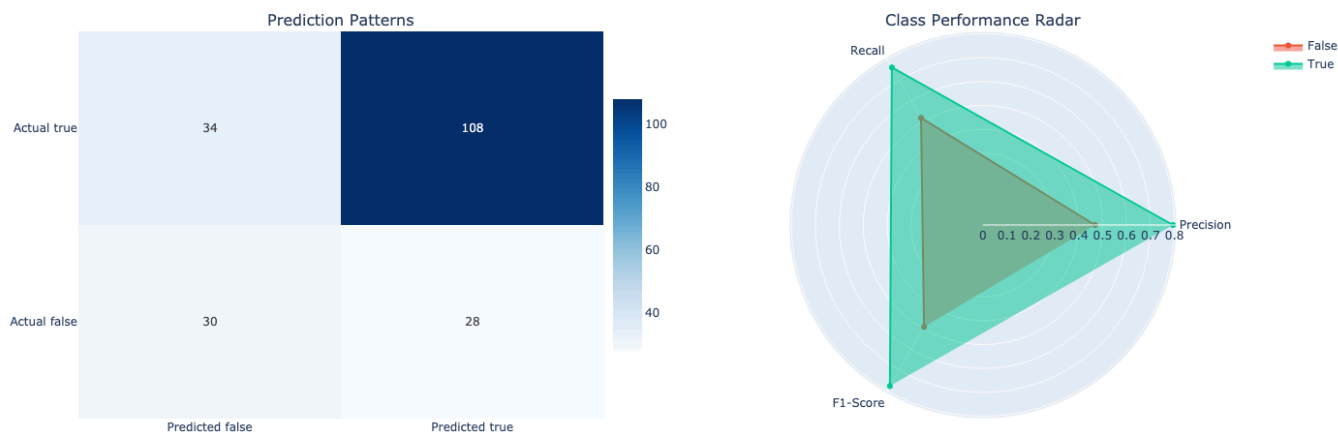
#### 2. Class Performance Radar (Recall, Precision, F1-Score) This radar chart visualizes model performance metrics across different classes:

- **Recall** → Measures how well the model identifies actual positives (high recall means fewer false negatives).
- **Precision** → Measures how many of the predicted positives were actually correct (high precision means fewer false positives).
- **F1-Score** → Balances precision and recall for a combined assessment.

#### What to look for & why:

- **Balanced scores across all three metrics** → Indicates a well-performing model with fair treatment across classes.
- **Low recall but high precision** → Means the model is very conservative, only predicting positives when it's very confident but missing many actual positives. This could be unfair if it leads to exclusion of certain groups.
- **High recall but low precision** → Suggests the model predicts positives too often, leading to many false positives, which can also be problematic in fairness-sensitive applications.
- **One class performing significantly worse** → Could indicate dataset imbalance, where the model learns patterns favoring the majority class.

💡 Takeaway: If performance varies significantly between classes, balancing the dataset or adjusting model thresholds can help improve fairness.



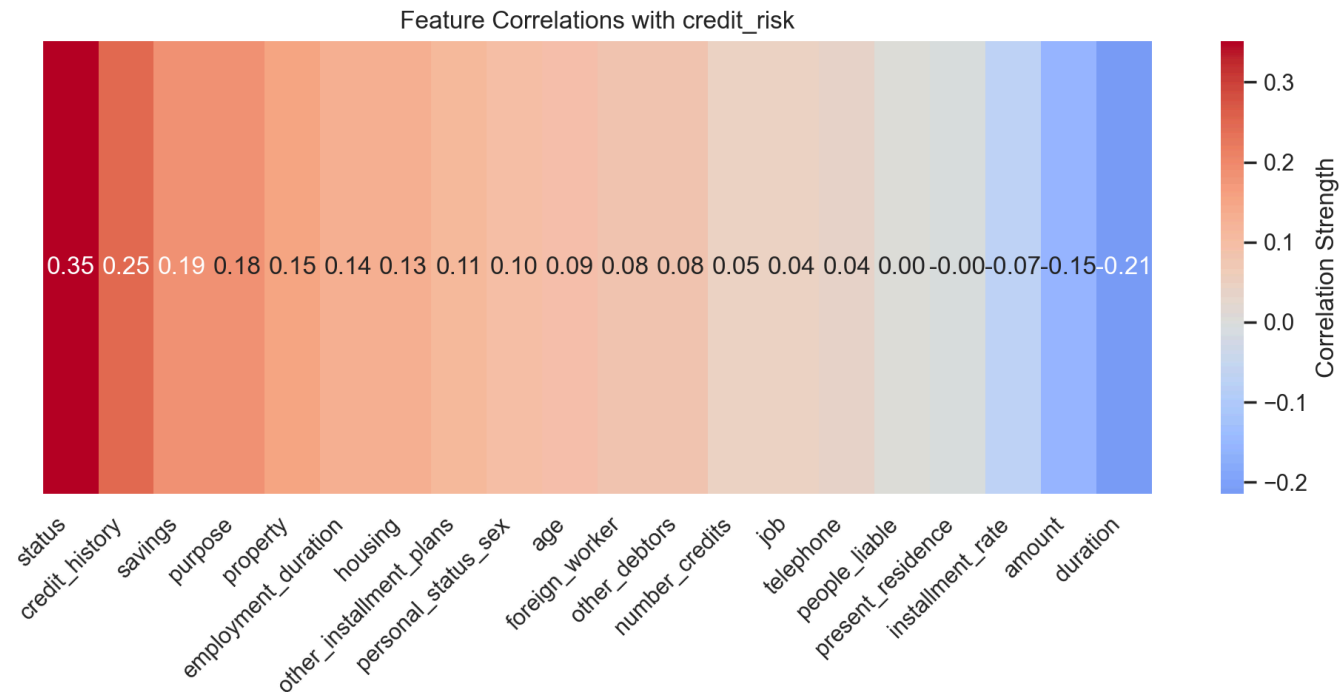
3. Feature Correlation (Prediction Attribute Impact) This heatmap shows how strongly each attribute correlates with the prediction outcome:

- **Blue colors** → Indicate weak or negative correlation.
- **Red colors** → Indicate strong positive correlation.
- **Darker shades** → Represent stronger correlation strength.

What to look for & why:

- **Features with extreme correlation (very red or very blue)** → May dominate predictions, potentially introducing bias. If a sensitive attribute (e.g., gender, race) has a high correlation, the model may be making unfair decisions based on it.
- **Low correlation for important attributes** → Suggests the model is not effectively using those features, possibly missing critical information.
- **Multiple highly correlated attributes** → Could indicate redundancy, where one feature indirectly represents another. This can cause fairness issues if an indirectly correlated sensitive feature influences decisions.

💡 Takeaway: If a sensitive attribute is highly correlated with the prediction, fairness interventions such as reweighting, debiasing techniques, or removing proxy variables may be necessary.



## Fairness Analysis & Protected Attribute Impact

These graphs focus on how fairness is distributed across protected attributes and whether the model exhibits bias toward specific groups.

### 1. Protected Attribute(s) Performance Breakdown

This line graph represents classification accuracy across different protected attribute groups.

- **Green areas indicate high accuracy** → The model performs well for those groups.
- **Red areas indicate lower accuracy** → The model struggles with classification for those groups.

What to look for & why:

- **Uneven accuracy across groups** → If accuracy varies significantly, some groups receive less reliable predictions, which can lead to fairness concerns.
- **Red-dominant sections** → Suggest that the model systematically underperforms for certain protected groups, potentially leading to discriminatory outcomes.
- **More green overall** → Indicates a more equitable model, meaning predictions are consistent across groups.

💡 Takeaway: If one group has significantly lower accuracy, consider balancing the dataset, adjusting model thresholds, or applying fairness-aware techniques to ensure fairer predictions.



True/False Positive/Negative rates by Protected Attribute



2. Top Correlation with Protected Values This boxplot visualizes how strongly each feature correlates with protected attributes (e.g., gender, race, age).

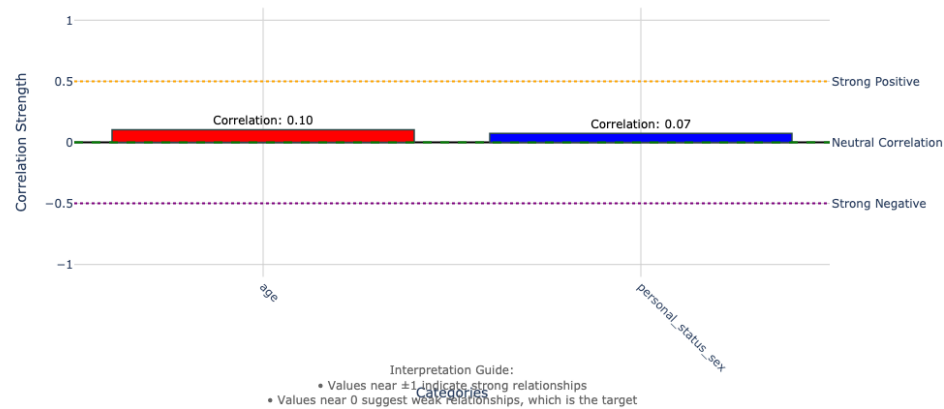
- **Values near  $\pm 1$**  → Indicate a strong correlation, meaning the feature is highly influenced by the protected attribute.
- **Values around 0** → Indicate weak correlation, meaning the feature is independent of the protected attribute.

What to look for & why:

- **Features with extreme correlation (close to -1 or +1)** → Suggest that protected attributes may be indirectly influencing predictions, which can introduce bias.
- **High correlation for non-sensitive attributes** → Could indicate proxy bias, where a seemingly neutral attribute is actually a disguised version of a protected feature.
- **Balanced, low correlations (closer to 0)** → Suggest a more fairness-aligned dataset, as predictions are less likely to be skewed by sensitive variables.

💡 Takeaway: If a protected attribute has a strong correlation with key features, consider removing proxy attributes, reweighting data, or using fairness-aware modeling techniques to mitigate bias.

## Top Correlations with Prediction Values



### 3. Fairness Analytics: Statistical Parity Difference This visualization evaluates fairness using Statistical Parity Difference (SPD):

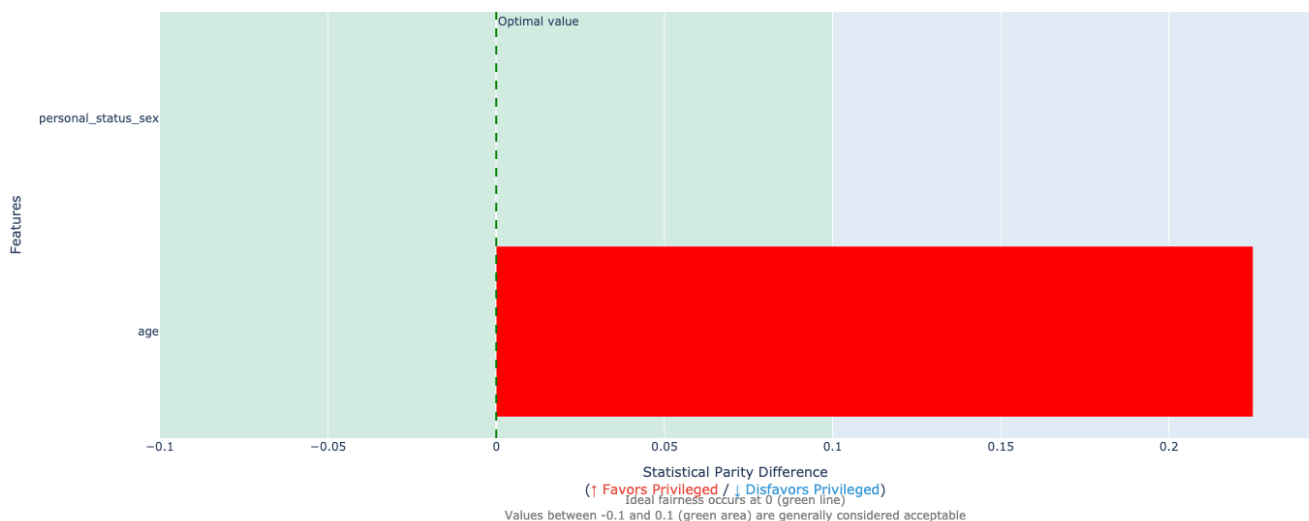
- **Attributes within the green threshold** → The model treats groups fairly, as the difference in selection rates between groups is within an acceptable range.
- **Attributes outside the threshold (red areas)** → Indicate bias, meaning certain groups are significantly more or less likely to receive positive predictions.

What to look for & why:

- **Green attributes (fair predictions)** → Suggests the model does not favor one group disproportionately.
- **Red attributes (bias detected)** → Indicate that a protected group is being disadvantaged or privileged unfairly. This may be due to dataset imbalance or model bias.
- **The further from the threshold, the more severe the bias** → Large deviations mean the model's decisions are disproportionately influenced by a sensitive attribute.

💡 Takeaway: If an attribute is flagged red, interventions like rebalancing the dataset, adjusting prediction thresholds, or applying fairness constraints may be necessary to reduce discrimination.




#### Fairness Analysis: Statistical Parity Differences




# Summary & Next Steps

This report provides a detailed fairness analysis of your dataset and model. By examining prediction patterns, class performance, feature correlations and fairness metrics, you can identify and mitigate biases before deploying your model.

## Key takeaways:

-  Balanced numerical and categorical distributions lead to more reliable models.
-  High correlation between sensitive attributes and predictions may indicate bias.
-  Unequal accuracy or statistical parity differences suggest fairness adjustments may be needed.

For best results, consider rebalancing data, adjusting thresholds, or applying fairness-aware techniques based on these insights.

 GitHub Repository (fork it): <https://github.com/Bence749/FairnessLens> Thank you for using FairnessLens—helping make AI more transparent, ethical, and fair! 