Analysis Report

Global dataset report

This report is the output of the Amazon SageMaker Clarify analysis. The report is split into following parts:

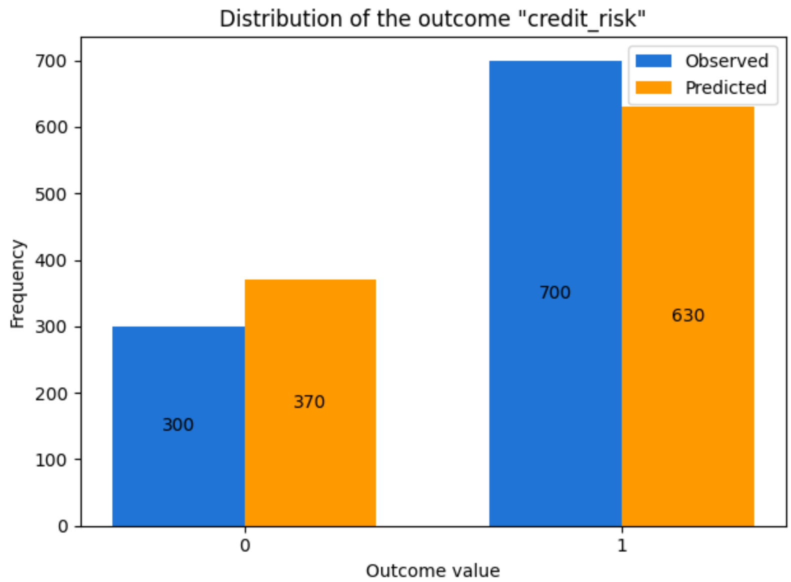
1. Analysis configuration
2. High level model performance
3. Pretraining bias metrics
4. Posttraining bias metrics

Analysis Configuration

Bias analysis requires you to configure the outcome label column, the facet and optionally a group variable. Generating explanations requires you to configure the outcome label. You configured the analysis with the following variables. The complete analysis configuration is appended at the end.

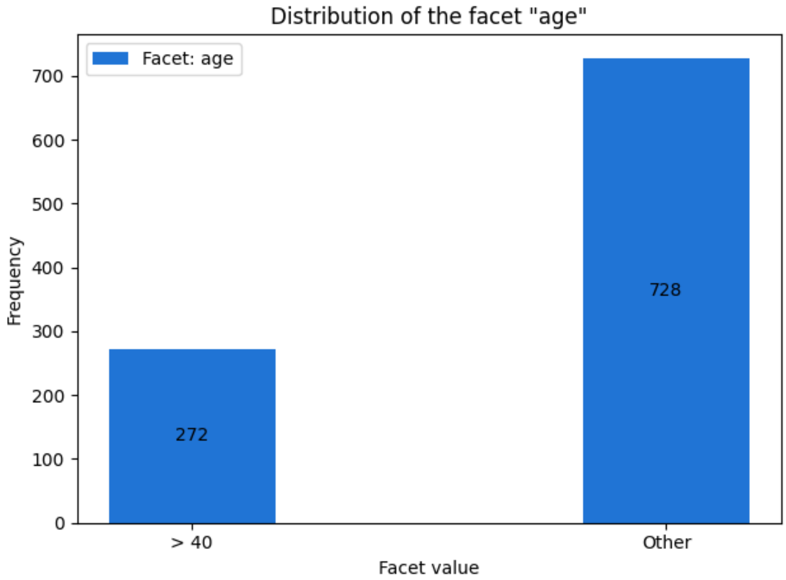
Outcome label: You chose the column `credit_risk` in the input data as the outcome label. Bias metric computation requires designating the positive outcome. You chose `credit_risk = 1` as the positive outcome. `credit_risk` consisted of values `[0, 1]` .

The figure below shows the distribution of values of `credit_risk` .



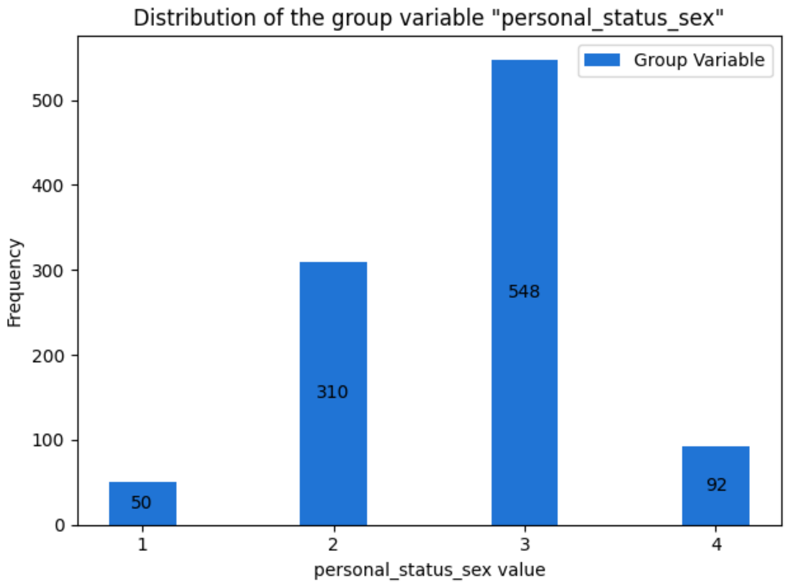
Facet: You chose the column `age` in the input data as the facet. `age` varied between `19.00` and `75.00` . Bias metrics were computed by comparing the inputs `age > 40` with all other inputs.

The figure below shows the distribution of values of `age` .



Group variable: Some bias metrics require an additional grouping variable. You chose `personal_status_sex` as the grouping variable.

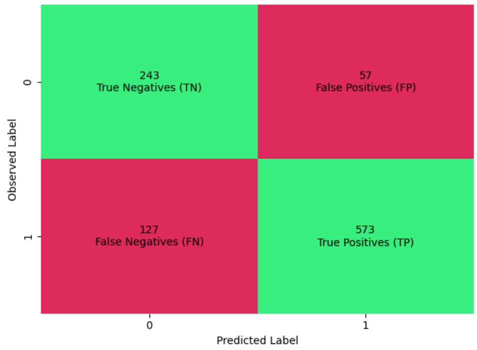
The figure below shows the distribution of values of `personal_status_sex`.



High level model performance

Input data points can be divided into different categories based on their observed and predicted label. For instance, a False Negative (FN) is an input with a positive observed label (`credit_risk = 1`) but negative predicted label (`credit_risk != 1`). A True Negative (TN) is an input whose observed and predicted labels are both negative.True Positives (TP) and False Positives (FP) are defined similarly.

Based on the model predictions, the inputs can be divided into different categories as:

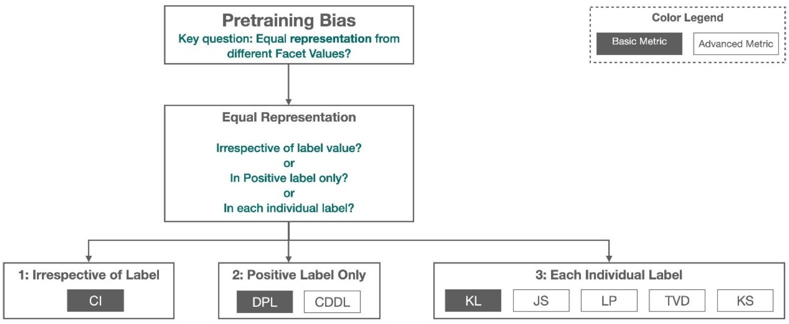


Here are metrics showing the model performance.

Metric	Description	Value
Accuracy	Proportion of inputs assigned the correct predicted label by the model.	0.816
Proportion of Positive Predictions in Labels	Proportion of input assigned in positive predicted label.	0.630
Proportion of Negative Predictions in Labels	Proportion of input assigned the negative predicted label.	0.370
True Positive Rate / Recall	Proportion of inputs with positive observed label correctly assigned the positive predicted label.	0.819
True Negative Rate / Specificity	Proportion of inputs with negative observed label correctly assigned the negative predicted label.	0.810
Acceptance Rate / Precision	Proportion of inputs with positive predicted label that actually have a positive observed label.	0.910
Rejection Rate	Proportion of inputs with negative predicted label that actually have a negative observed label.	0.657
Conditional Acceptance	Ratio between the positive observed labels and positive predicted labels.	1.111
Conditional Rejection	Ratio between the negative observed labels and negative predicted labels.	0.811
F1 Score	Harmonic mean of precision and recall.	0.862

Pre-training Bias Metrics

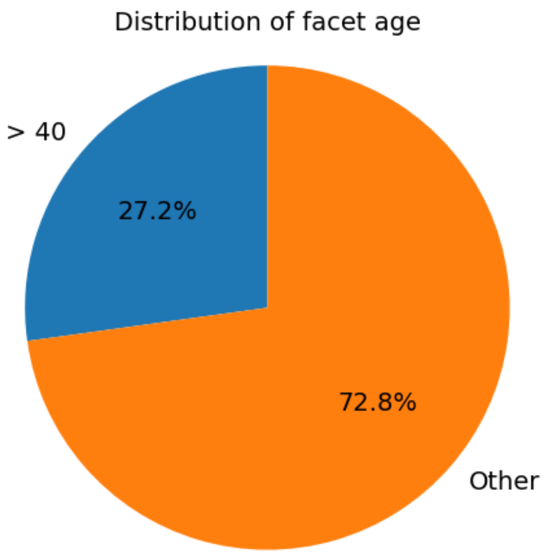
Pretraining bias metrics measure imbalances in facet value representation in the training data. Imbalances can be measured across different dimensions. For instance, you could focus imbalances within the inputs with positive observed label only. The figure below shows how different pretraining bias metrics focus on different dimensions. For a detailed description of these dimensions, see [Learn How Amazon SageMaker Clarify Helps Detect Bias](#).



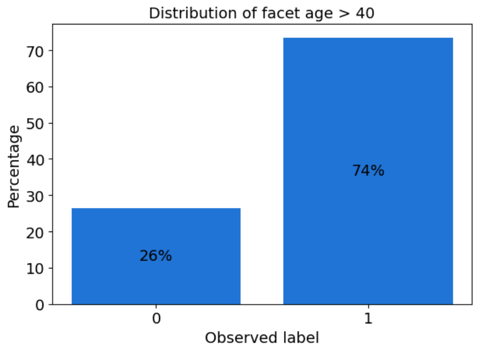
The metric values along with an informal description of what they mean are shown below. For mathematical formulas and examples, see the [Measure Pretraining Bias](#) section of the AWS documentation.

We computed the bias metrics for the label `credit_risk` using label value(s)/threshold `credit_risk = 1` for the following facets:

- Facet column: **age**
  The pie chart shows the distribution of facet column `age` in your data.



The bar plot(s) below show the distribution of facet column `age` in your data.

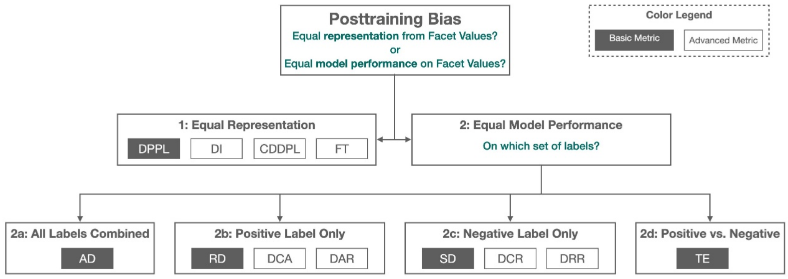


Facet Value(s)/Threshold: `age > 40`

Metric	Description	Value
[Conditional Demographic Disparity in Labels (CDDL)]()	Measures maximum divergence between the observed label distributions for facet values `age > 40` and rest of the inputs in the dataset.	-0.035
[Class Imbalance (CI)]()	Measures the imbalance in the number of inputs with facet values `age > 40` and rest of the inputs.	0.456
[Difference in Proportions of Labels (DPL)]()	Measures the imbalance of positive observed labels between facet values `age > 40` and rest of the inputs.	-0.048
[Jensen-Shannon Divergence (JS)]()	Measures how much the observed label distributions of facet values `age > 40` and rest of the inputs diverge from each other entropically.	0.001
[Kullback-Leibler Divergence (KL)]()	Measures how much the observed label distributions of facet values `age > 40` and rest of the inputs diverge from each other entropically.	0.006
[Kolmogorov-Smirnov (KS)]()	Measures maximum divergence between the observed label distributions for facet values `age > 40` and rest of the inputs in the dataset.	0.048
[Lp-norm (LP)]()	Measures a p-norm difference between the observed label distributions associated with facet values `age > 40` rest of the inputs in the dataset.	0.069
[Total Variation Distance (TVD)]()	Measures half of the L1-norm difference between the observed label distributions associated with facet values `age > 40` and rest of the inputs in the dataset.	0.048

Post-training Bias Metrics

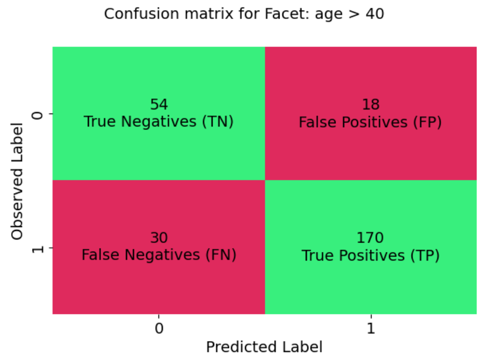
Posttraining bias metrics measure imbalances in model predictions across different inputs. The figure below shows how different posttraining metrics target different types of imbalances over inputs. For a detailed description of these types, see [Learn How Amazon SageMaker Clarify Helps Detect Bias]().



Bias can also result form imbalances in the model outcomes even when the facet value is not considered. The metric computing these imbalances is GE. The metric values along with an informal description of what they mean are shown below. For mathematical formulas and examples, see the Measure Posttraining Data and Model Bias section of the AWS documentation.

We computed the bias metrics for the label `credit_risk` using label value(s)/threshold `credit_risk = 1` for the following facets:

- Facet column: **age**
  Facet Value(s)/Threshold: `age > 40`



Metric	Description	Value
Accuracy Difference (AD)	Measures the difference between the prediction accuracy for facet values `age > 40` and rest of the inputs.	-0.010
Conditional Demographic Disparity in Predicted Labels (CDDPL)	Measures the disparity of predicted labels between facet values `age > 40` and rest of the inputs as a whole, but also by subgroups dictated by Age.	-0.062
Difference in Acceptance Rates (DAR)	Measures the difference in the ratios of the observed positive outcomes (TP) to the predicted positives (TP + FP) between facet values `age > 40` and rest of the inputs.	0.008
Difference in Conditional Acceptance (DCAcc)	Compares the observed labels to the labels predicted by the model. Assesses whether this is the same across facet values `age > 40` and rest of the inputs for predicted positive outcomes (acceptances).	0.067
Difference in Conditional Rejection (DCR)	Compares the observed labels to the labels predicted by the model and assesses whether this is the same across facet values `age > 40` and rest of the inputs for negative outcomes (rejections).	0.060
Disparate Impact (DI)	Measures the ratio of proportions of the predicted labels for facet values `age > 40` and rest of the inputs.	1.138
Difference in Positive Proportions in Predicted Labels (DPPL)	Measures the difference in the proportion of positive predictions between facet values `age > 40` and rest of the inputs.	-0.084
Difference in Rejection Rates (DRR)	Measures the difference in the ratios of the observed negative outcomes (TN) to the predicted negatives (TN + FN) between facet values `age > 40` and rest of the inputs.	-0.018
Counterfactual Fliptest (FT)	Examines each input with facet value `age > 40` and assesses whether similar members from rest of the inputs have different model predictions.	0.066
Generalized entropy (GE)	Measures the inequality in benefits b assigned to each input by the model predictions.	0.104
Recall Difference (RD)	Measures the difference between the recall, aka true positive rate, of the model for facet values `age > 40` and rest of the inputs.	-0.044
Specificity difference (SD)	Measures the difference between the specificity, aka true negative rate, of the model for facet values `age > 40` and rest of the inputs.	-0.079
Treatment Equality (TE)	Measures the difference in the ratio of false positives to false negatives between facet values `age > 40` and rest of the inputs.	-0.821

Appendix: Analysis Configuration Parameters


```
{
  "dataset_type": "text/csv",
  "headers": [
    "status",
    "duration",
    "credit_history",
    "purpose",
    "amount",
    "savings",
    "employment_duration",
    "installment_rate",
    "personal_status_sex",
    "other_debtors",
    "present_residence",
    "property",
    "age",
    "other_installment_plans",
    "housing",
    "number_credits",
    "job",
    "people_liable",
    "telephone",
    "foreign_worker",
    "credit_risk"
  ],
  "label": "credit_risk",
  "label_values_or_threshold": [
    1
  ],
  "facet": [
    {
      "name_or_index": "age",
      "value_or_threshold": [
        40
      ]
    }
  ],
  "group_variable": "personal_status_sex",
  "methods": {
    "report": {
      "name": "report",
      "title": "Analysis Report"
    },
    "pre_training_bias": {
      "methods": "all"
    },
    "post_training_bias": {
      "methods": "all"
    }
  },
  "predictor": {
    "model_name": "credit-risk-inference-pipeline-1705694102",
    "instance_type": "ml.c5.xlarge",
    "initial_instance_count": 1,
    "accept_type": "text/csv"
  },
  "probability_threshold": 0.7
}
```