Model Monitor Bias Report

Global dataset report

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

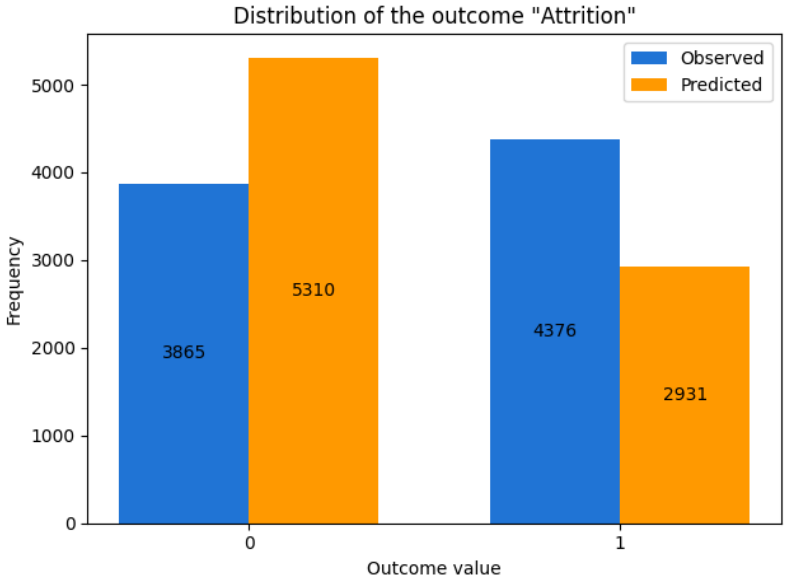
1. Analysis configuration
2. High level model performance
3. Pretraining bias metrics
4. Posttraining bias metrics

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.

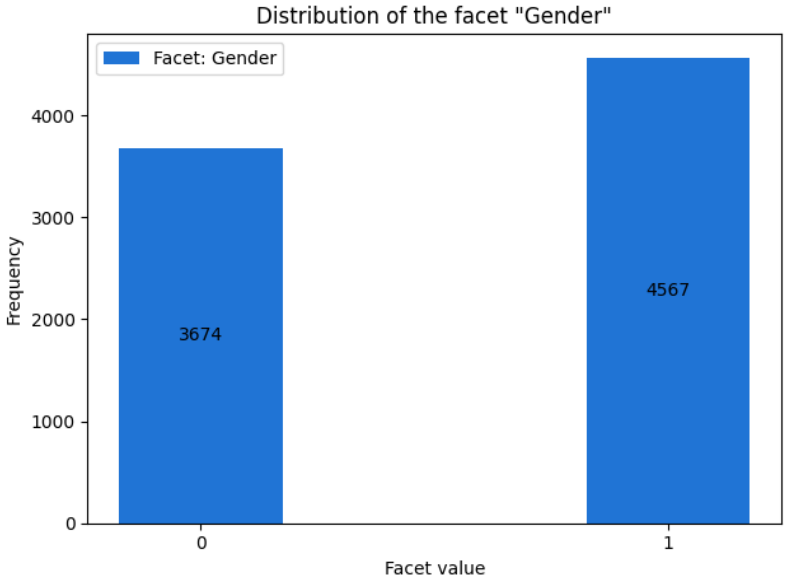
Outcome label: You chose the column `Attrition` in the input data as the outcome label. Bias metric computation requires designating the positive outcome. You chose `Attrition = 1` as the positive outcome. `Attrition` consisted of values `[0, 1]`.

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



Facet: You chose the column `Gender` in the input data as the facet. `Gender` consisted of values `[0, 1]`. Bias metrics were computed by comparing the inputs `Gender = 0,1` with all other inputs.

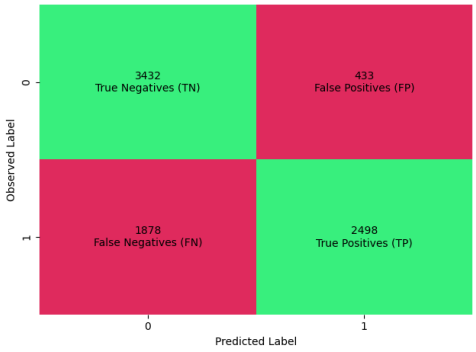
The figure below shows the distribution of values of `Gender`.



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 (`Attrition = 1`) but negative predicted label (`Attrition != 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:

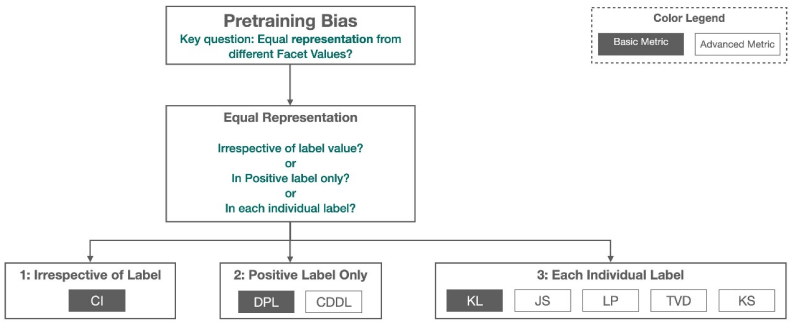


Here are metrics showing the model performance.

Metric	Description	Value
Accuracy	Proportion of inputs assigned the correct predicted label by the model.	0.720
Proportion of Positive Predictions in Labels	Proportion of input assigned in positive predicted label.	0.356
Proportion of Negative Predictions in Labels	Proportion of input assigned the negative predicted label.	0.644
True Positive Rate / Recall	Proportion of inputs with positive observed label correctly assigned the positive predicted label.	0.571
True Negative Rate / Specificity	Proportion of inputs with negative observed label correctly assigned the negative predicted label.	0.888
Acceptance Rate / Precision	Proportion of inputs with positive predicted label that actually have a positive observed label.	0.852
Rejection Rate	Proportion of inputs with negative predicted label that actually have a negative observed label.	0.646
Conditional Acceptance	Ratio between the positive observed labels and positive predicted labels.	1.493
Conditional Rejection	Ratio between the negative observed labels and negative predicted labels.	0.728
F1 Score	Harmonic mean of precision and recall.	0.684

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.

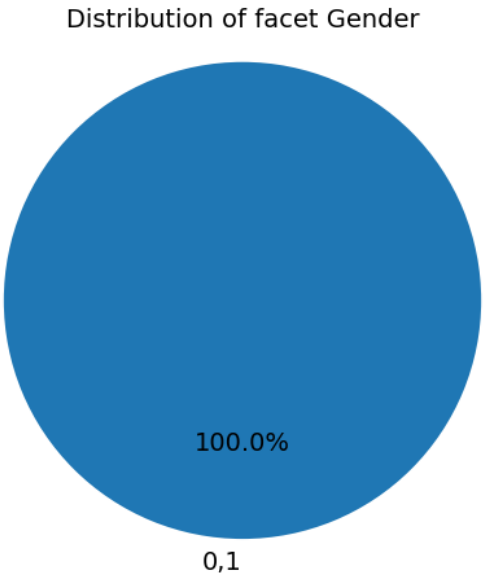


The metric values along with an informal description of what they mean are shown below. For mathematical formulas and examples, see the [Measure Pretraining Bias](https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-measure-data-bias.html) section of the AWS documentation.

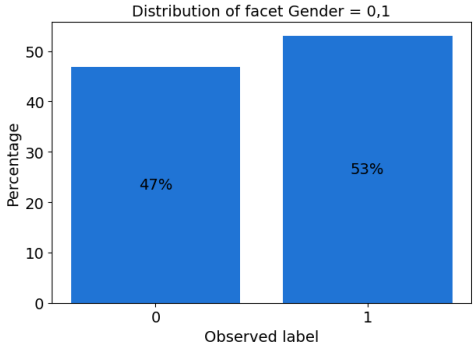
We computed the bias metrics for the label `Attrition` using label value(s)/threshold `Attrition = 1` for the following facets:

- Facet column: **Gender**

The pie chart shows the distribution of facet column `Gender` in your data.



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

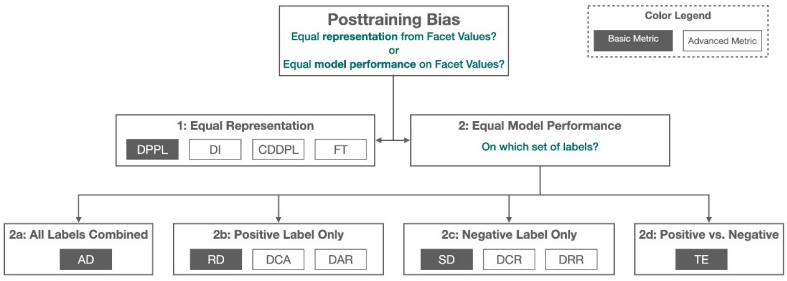


Facet Value(s)/Threshold: `Gender = 0,1`

Metric	Description	Value	Error
Conditional Demographic Disparity in Labels (CDDL)	Measures maximum divergence between the observed label distributions for facet values `Gender = 0,1` and rest of the inputs in the dataset.	None	Error: see Clarify job output
Class Imbalance (CI)	Measures the imbalance in the number of inputs with facet values `Gender = 0,1` and rest of the inputs.	None	Error: see Clarify job output
Difference in Proportions of Labels (DPL)	Measures the imbalance of positive observed labels between facet values `Gender = 0,1` and rest of the inputs.	None	Error: see Clarify job output
Jensen-Shannon Divergence (JS)	Measures how much the observed label distributions of facet values `Gender = 0,1` and rest of the inputs diverge from each other entropically.	None	Error: see Clarify job output
Kullback-Leibler Divergence (KL)	Measures how much the observed label distributions of facet values `Gender = 0,1` and rest of the inputs diverge from each other entropically.	None	Error: see Clarify job output
Kolmogorov-Smirnov (KS)	Measures maximum divergence between the observed label distributions for facet values `Gender = 0,1` and rest of the inputs in the dataset.	None	Error: see Clarify job output
Lp-norm (LP)	Measures a p-norm difference between the observed label distributions associated with facet values `Gender = 0,1` rest of the inputs in the dataset.	None	Error: see Clarify job output
Total Variation Distance (TVD)	Measures half of the L1-norm difference between the observed label distributions associated with facet values `Gender = 0,1` and rest of the inputs in the dataset.	None	Error: see Clarify job output

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.

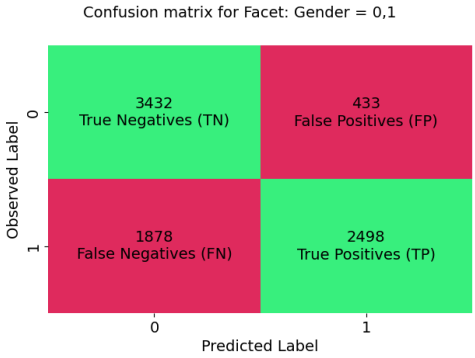


Bias can also result form imbalances in the model outcomes even when the facet value is not considered. The metric computing these imbalances is GE. The metric values along with an informal description of what they mean are shown below. For mathematical formulas and examples, see the [Measure Posttraining Data and Model Bias] (https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-measure-post-training-bias.html) section of the AWS documentation.

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

- Facet column: **Gender**

Facet Value(s)/Threshold: `Gender = 0,1`


Confusion matrix for Facet: Gender = 0,1
3432
True Negatives (TN)
433
False Positives (FP)
1878
False Negatives (FN)
2498
True Positives (TP)
Observed Label
0
1
0
1
Predicted Label

Metric	Description	Value	Error
Accuracy Difference (AD)	Measures the difference between the prediction accuracy for facet values `Gender = 0,1` and rest of the inputs.	None	Error: see Clarify job output
Conditional Demographic Disparity in Predicted Labels (CDDPL)	Measures the disparity of predicted labels between facet values `Gender = 0,1` and rest of the inputs as a whole, but also by subgroups dictated by Age.	None	Error: see Clarify job output
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 `Gender = 0,1` and rest of the inputs.	None	Error: see Clarify job output
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 `Gender = 0,1` and rest of the inputs for predicted positive outcomes (acceptances).	None	Error: see Clarify job output
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 `Gender = 0,1` and rest of the inputs for negative outcomes (rejections).	None	Error: see Clarify job output
Disparate Impact (DI)	Measures the ratio of proportions of the predicted labels for facet values `Gender = 0,1` and rest of the inputs.	None	Error: see Clarify job output
Difference in Positive Proportions in Predicted Labels (DPPL)	Measures the difference in the proportion of positive predictions between facet values `Gender = 0,1` and rest of the inputs.	None	Error: see Clarify job output
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 `Gender = 0,1` and rest of the inputs.	None	Error: see Clarify job output
Counterfactual Fliptest (FT)	Examines each input with facet value `Gender = 0,1` and assesses whether similar members from rest of the inputs have different model predictions.	None	Error: see Clarify job output
Generalized entropy (GE)	Measures the inequality in benefits b assigned to each input by the model predictions.	0.184	None
Recall Difference (RD)	Measures the difference between the recall, aka true positive rate, of the model for facet values `Gender = 0,1` and rest of the inputs.	None	Error: see Clarify job output
Specificity difference (SD)	Measures the difference between the specificity, aka true negative rate, of the model for facet values `Gender = 0,1` and rest of the inputs.	None	Error: see Clarify job output
Treatment Equality (TE)	Measures the difference in the ratio of false positives to false negatives between facet values `Gender = 0,1` and rest of the inputs.	None	Error: see Clarify job output

Appendix: Analysis Configuration Parameters


```
{
   "label_values_or_threshold": [
      1
   ],
   "facet": [
```

```
        {
            "name_or_index": "Gender",
            "value_or_threshold": [
                0,
                1
            ]
        }
    ],
    "headers": [
        "Employee ID",
        "Age",
        "Gender",
        "Years at Company",
        "Job Role",
        "Monthly Income",
        "Work-Life Balance",
        "Job Satisfaction",
        "Performance Rating",
        "Number of Promotions",
        "Overtime",
        "Distance from Home",
        "Education Level",
        "Marital Status",
        "Number of Dependents",
        "Job Level",
        "Company Size",
        "Company Tenure",
        "Remote Work",
        "Leadership Opportunities",
        "Innovation Opportunities",
        "Company Reputation",
        "Employee Recognition",
        "Attrition"
    ],
    "label": "Attrition",
    "dataset_type": "application/sagemakercapturejson",
    "methods": {
        "pre_training_bias": {
            "methods": "all"
        },
        "post_training_bias": {
            "methods": "all"
        },
        "report": {
            "name": "report",
            "title": "Model Monitor Bias Report"
        }
    },
    "probability_threshold": 0.7,
    "predictor": {}
}
```