Analysis Report

Global dataset report

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

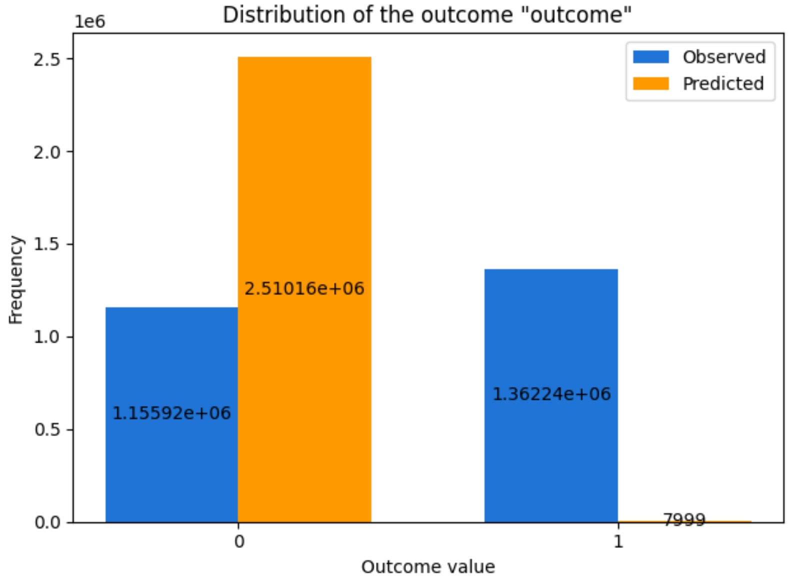
1. Analysis configuration
2. High level model performance
3. Posttraining bias metrics

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.

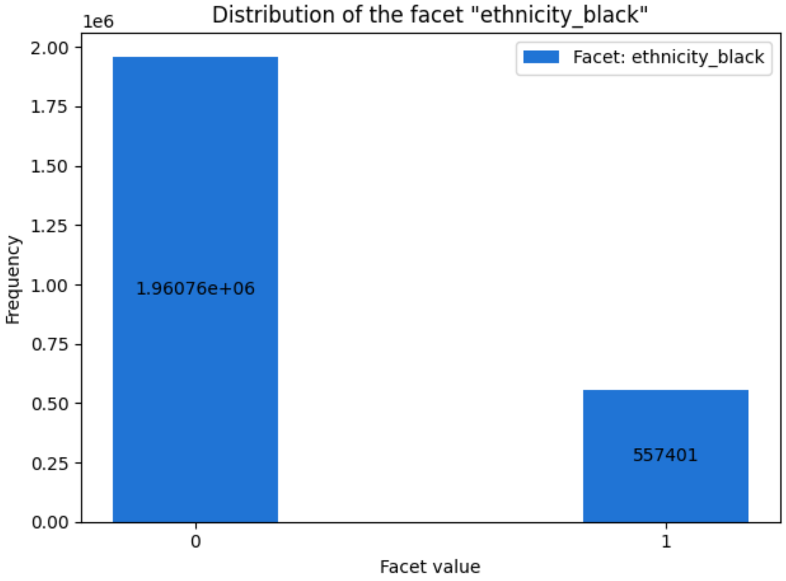
Outcome label: You chose the column `outcome` in the input data as the outcome label. Bias metric computation requires designating the positive outcome. You chose `outcome = 0` as the positive outcome. `outcome` consisted of values `[0, 1]` .

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



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

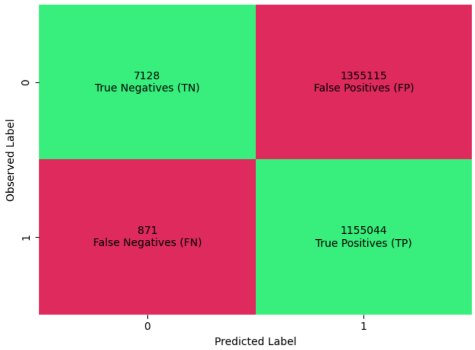
The figure below shows the distribution of values of `ethnicity_black` .



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 (`outcome = 0`) but negative predicted label (`outcome != 0`). 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:

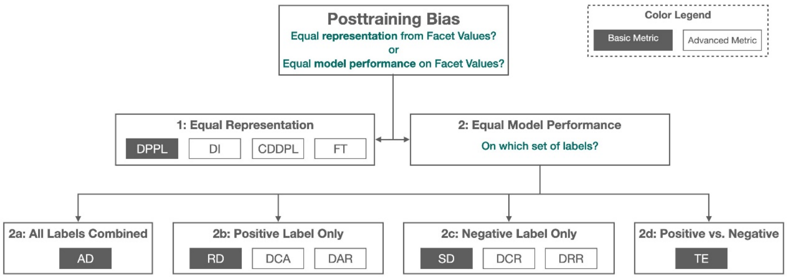


Here are metrics showing the model performance.

Metric	Description	Value
Accuracy	Proportion of inputs assigned the correct predicted label by the model.	0.462
Proportion of Positive Predictions in Labels	Proportion of input assigned in positive predicted label.	0.997
Proportion of Negative Predictions in Labels	Proportion of input assigned the negative predicted label.	0.003
True Positive Rate / Recall	Proportion of inputs with positive observed label correctly assigned the positive predicted label.	0.999
True Negative Rate / Specificity	Proportion of inputs with negative observed label correctly assigned the negative predicted label.	0.005
Acceptance Rate / Precision	Proportion of inputs with positive predicted label that actually have a positive observed label.	0.460
Rejection Rate	Proportion of inputs with negative predicted label that actually have a negative observed label.	0.891
Conditional Acceptance	Ratio between the positive observed labels and positive predicted labels.	0.460
Conditional Rejection	Ratio between the negative observed labels and negative predicted labels.	170.302
F1 Score	Harmonic mean of precision and recall.	0.630

Post-training Bias Metrics

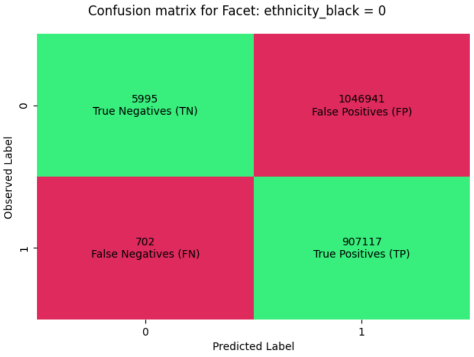
Posttraining bias metrics measure imbalances in model predictions across different inputs. The figure below shows how different posttraining metrics target different types of imbalances over inputs. For a detailed description of these types, see Learn How Amazon SageMaker Clarify Helps Detect Bias.



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.

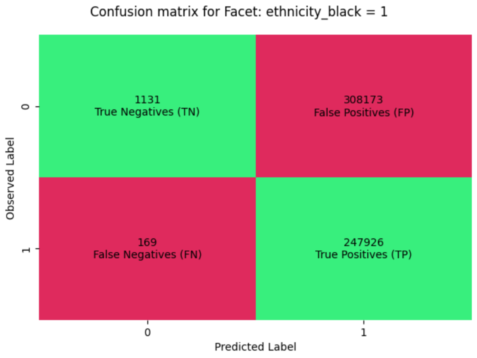
We computed the bias metrics for the label `outcome` using label value(s)/threshold `outcome = 0` for the following facets:

- Facet column: **ethnicity_black**
  Facet Value(s)/Threshold: ethnicity_black = 0



Metric	Description	Value	Error
Accuracy Difference (AD)	Measures the difference between the prediction accuracy for facet values Sex=0 and rest of the inputs.	-0.019	None
Conditional Demographic Disparity in Predicted Labels (CDDPL)	Measures the disparity of predicted labels between facet values Sex=0 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 Sex=0 and rest of the inputs.	-0.018	None
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 Sex=0 and rest of the inputs for predicted positive outcomes (acceptances).	-0.018	None
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 Sex=0 and rest of the inputs for negative outcomes (rejections).	-80.543	None
Disparate Impact (DI)	Measures the ratio of proportions of the predicted labels for facet values Sex=0 and rest of the inputs.	0.999	None
Difference in Positive Proportions in Predicted Labels (DPPL)	Measures the difference in the proportion of positive predictions between facet values Sex=0 and rest of the inputs.	0.001	None
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 Sex=0 and rest of the inputs.	0.025	None
Counterfactual Fliptest (FT)	Examines each input with facet value Sex=0 and assesses whether similar members from rest of the inputs have different model predictions.	-0.002	None
Generalized entropy (GE)	Measures the inequality in benefits b assigned to each input by the model predictions.	0.053	None
Recall Difference (RD)	Measures the difference between the recall, aka true positive rate, of the model for facet values Sex=0 and rest of the inputs.	0.000	None
Specificity difference (SD)	Measures the difference between the specificity, aka true negative rate, of the model for facet values Sex=0 and rest of the inputs.	0.002	None
Treatment Equality (TE)	Measures the difference in the ratio of false positives to false negatives between facet values Sex=0 and rest of the inputs.	0.000	None

Facet Value(s)/Threshold: ethnicity_black = 1



Metric	Description	Value	Error
Accuracy Difference (AD)	Measures the difference between the prediction accuracy for facet values Sex=0 and rest of the inputs.	0.019	None
Conditional Demographic Disparity in Predicted Labels (CDDPL)	Measures the disparity of predicted labels between facet values Sex=0 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 Sex=0 and rest of the inputs.	0.018	None
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 Sex=0 and rest of the inputs for predicted positive outcomes (acceptances).	0.018	None
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 Sex=0 and rest of the inputs for negative outcomes (rejections).	80.543	None
Disparate Impact (DI)	Measures the ratio of proportions of the predicted labels for facet values Sex=0 and rest of the inputs.	1.001	None
Difference in Positive Proportions in Predicted Labels (DPPL)	Measures the difference in the proportion of positive predictions between facet values Sex=0 and rest of the inputs.	-0.001	None
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 Sex=0 and rest of the inputs.	-0.025	None
Counterfactual Fliptest (FT)	Examines each input with facet value Sex=0 and assesses whether similar members from rest of the inputs have different model predictions.	-0.001	None
Generalized entropy (GE)	Measures the inequality in benefits b assigned to each input by the model predictions.	0.053	None
Recall Difference (RD)	Measures the difference between the recall, aka true positive rate, of the model for facet values Sex=0 and rest of the inputs.	-0.000	None
Specificity difference (SD)	Measures the difference between the specificity, aka true negative rate, of the model for facet values Sex=0 and rest of the inputs.	-0.002	None
Treatment Equality (TE)	Measures the difference in the ratio of false positives to false negatives between facet values Sex=0 and rest of the inputs.	-0.000	None

Appendix: Analysis Configuration Parameters

```
{
    "dataset_type": "text/csv",
    "headers": [
        "outcome",
        "gender_male",
        "gender_female",
        "gender_other",
        "age_range_18-24",
        "age_range_25-34",
        "age_range_over_34",
        "age_range_10-17",
        "age_range_under_10",
        "date_month_1",
        "date_month_2",
        "date_day_1",
        "date_day_2",
        "date_hour_1",
        "date_hour_2",
        "date_week_of_year_1",
        "date_week_of_year_2",
        "date_day_of_year_1",
        "date_day_of_year_2",
        "date_quarter_1",
        "date_quarter_2",
        "ethnicity_white",
        "ethnicity_black",
    "ethnicity_asian",
    "ethnicity_other",
    "ethnicity_mixed",
    "coords"
  ],
  "label": "outcome",
  "label_values_or_threshold": [
    0
  ],
  "facet": [
    {
      "name_or_index": "ethnicity_black"
    }
  ],
  "methods": {
    "report": {
      "name": "report",
      "title": "Analysis Report"
    },
    "post_training_bias": {
      "methods": "all"
    }
  },
  "predictor": {
    "endpoint_name": "xgboost-2023-04-01-11-26-14-202",
    "accept_type": "text/csv",
    "content_type": "text/csv"
  },
  "probability_threshold": 0.8
}
```