Analysis Report

Global dataset report

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

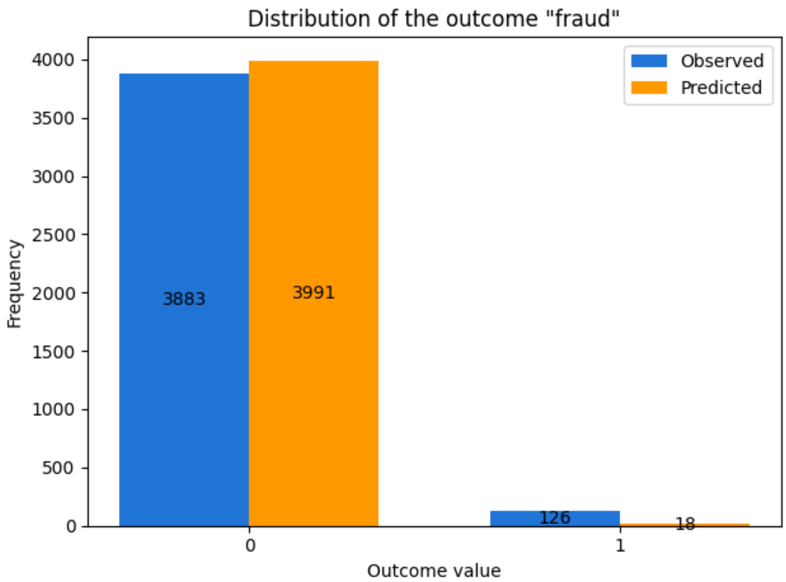
1. Analysis configuration
2. High level model performance
3. Pretraining bias metrics
4. Posttraining bias metrics

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.

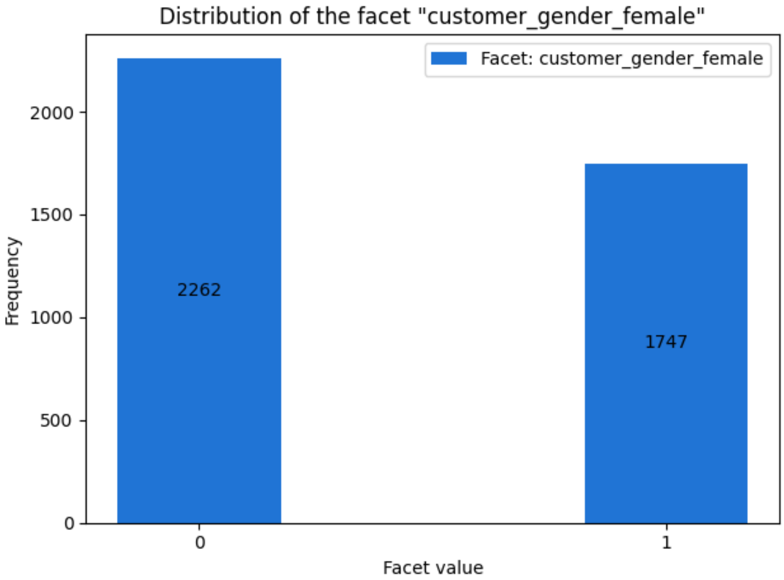
Outcome label: You chose the column `fraud` in the input data as the outcome label. Bias metric computation requires designating the positive outcome. You chose `fraud = 0` as the positive outcome. `fraud` consisted of values `[0, 1]`.

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



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

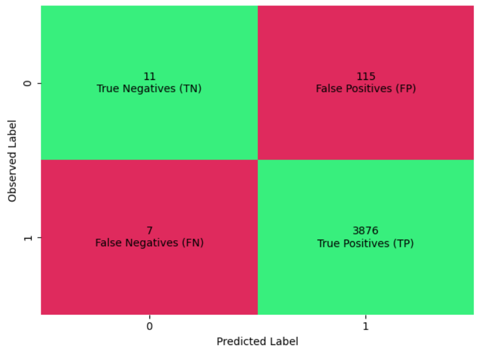
The figure below shows the distribution of values of `customer_gender_female`.



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 (fraud = 0) but negative predicted label (fraud != 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:

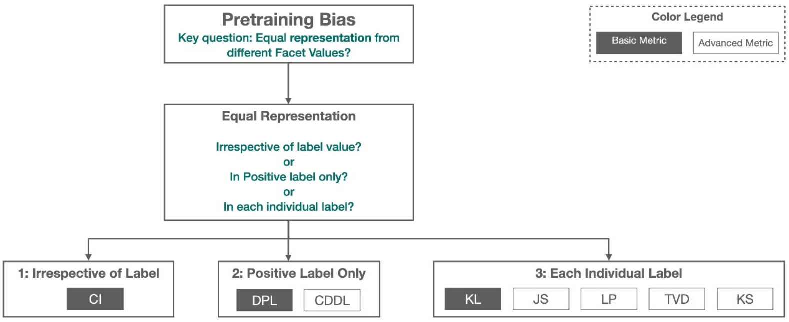


Here are metrics showing the model performance.

Metric	Description	Value
Accuracy	Proportion of inputs assigned the correct predicted label by the model.	0.970
Proportion of Positive Predictions in Labels	Proportion of input assigned in positive predicted label.	0.996
Proportion of Negative Predictions in Labels	Proportion of input assigned the negative predicted label.	0.004
True Positive Rate / Recall	Proportion of inputs with positive observed label correctly assigned the positive predicted label.	0.998
True Negative Rate / Specificity	Proportion of inputs with negative observed label correctly assigned the negative predicted label.	0.087
Acceptance Rate / Precision	Proportion of inputs with positive predicted label that actually have a positive observed label.	0.971
Rejection Rate	Proportion of inputs with negative predicted label that actually have a negative observed label.	0.611
Conditional Acceptance	Ratio between the positive observed labels and positive predicted labels.	0.973
Conditional Rejection	Ratio between the negative observed labels and negative predicted labels.	7.000
F1 Score	Harmonic mean of precision and recall.	0.985

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](https://aws.amazon.com/blogs/machine-learning/learn-how-amazon-sagemaker-clarify-helps-detect-bias).

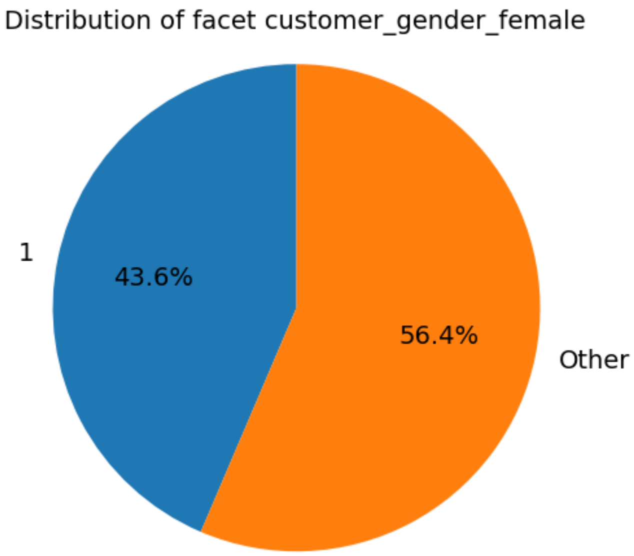


The metric values along with an informal description of what they mean are shown below. For mathematical formulas and examples, see the [Measure Pretraining Bias](https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-measure-data-bias.html) section of the AWS documentation.

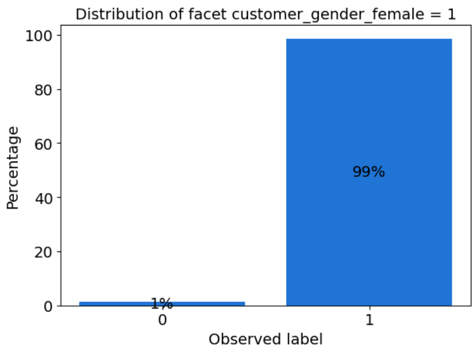
We computed the bias metrics for the label `fraud` using label value(s)/threshold `fraud = 0` for the following facets:

- Facet column: **customer_gender_female**

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



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

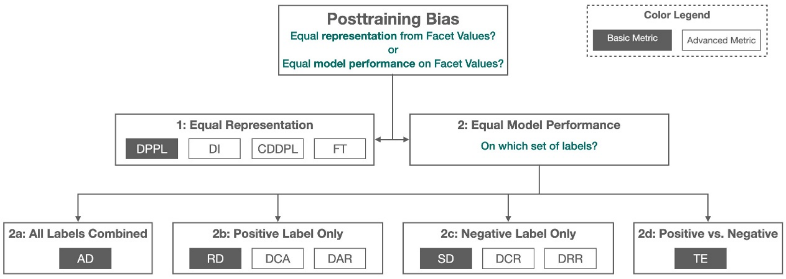


Facet Value(s)/Threshold: `customer_gender_female = 1`

Metric	Description	Value
[Class Imbalance (CI)]	Measures the imbalance in the number of inputs with facet values `customer_gender_female = 1` and rest of the inputs.	0.128

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].

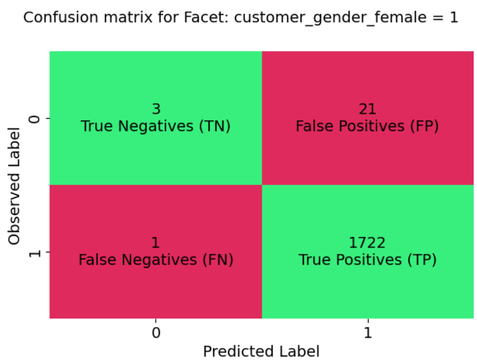


Bias can also result form imbalances in the model outcomes even when the facet value is not considered. The metric computing these imbalances is GE. The metric values along with an informal description of what they mean are shown below. For mathematical formulas and examples, see the [Measure Posttraining 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 `fraud` using label value(s)/threshold `fraud = 0` for the following facets:

- Facet column: **customer_gender_female**

Facet Value(s)/Threshold: `customer_gender_female = 1`



Metric	Description	Value
[Difference in Positive Proportions in Predicted Labels (DPPL)]	Measures the difference in the proportion of positive predictions between facet values `customer_gender_female = 1` and rest of the inputs.	-0.004

Appendix: Analysis Configuration Parameters


```
{
  "dataset_type": "text/csv",
  "headers": [
    "fraud",
    "num_vehicles_involved",
    "num_injuries",
    "num_witnesses",
```

```
    "police_report_available",
    "injury_claim",
    "vehicle_claim",
    "total_claim_amount",
    "incident_month",
    "incident_day",
    "incident_dow",
    "incident_hour",
    "customer_age",
    "months_as_customer",
    "num_claims_past_year",
    "num_insurers_past_5_years",
    "policy_deductable",
    "policy_annual_premium",
    "policy_liability",
    "customer_education",
    "auto_year",
    "driver_relationship_other",
    "driver_relationship_child",
    "driver_relationship_spouse",
    "driver_relationship_na",
    "driver_relationship_self",
    "incident_type_collision",
    "incident_type_break-in",
    "incident_type_theft",
    "collision_type_rear",
    "collision_type_side",
    "collision_type_na",
    "collision_type_front",
    "incident_severity_totaled",
    "incident_severity_major",
    "incident_severity_minor",
    "authorities_contacted_fire",
    "authorities_contacted_none",
    "authorities_contacted_police",
    "authorities_contacted_ambulance",
    "policy_state_ca",
    "policy_state_az",
    "policy_state_nv",
    "policy_state_id",
    "policy_state_wa",
    "policy_state_or",
    "customer_gender_other",
    "customer_gender_male",
    "customer_gender_female"
  ],
  "label": "fraud",
  "label_values_or_threshold": [
    0
  ],
  "facet": [
    {
      "name_or_index": "customer_gender_female",
      "value_or_threshold": [
        1
      ]
    }
  ],
  "methods": {
    "report": {
      "name": "report",
      "title": "Analysis Report"
    },
    "pre_training_bias": {
      "methods": [
        "CI"
```

```
                CI
            ]
        },
        "post_training_bias": {
            "methods": [
                "DPPL"
            ]
        }
    },
    "predictor": {
        "model_name": "fraud-detect-xgb-model",
        "instance_type": "ml.m5.xlarge",
        "initial_instance_count": 1,
        "accept_type": "text/csv"
    },
    "probability_threshold": 0.5
}
```