Analysis Report

Global dataset report

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

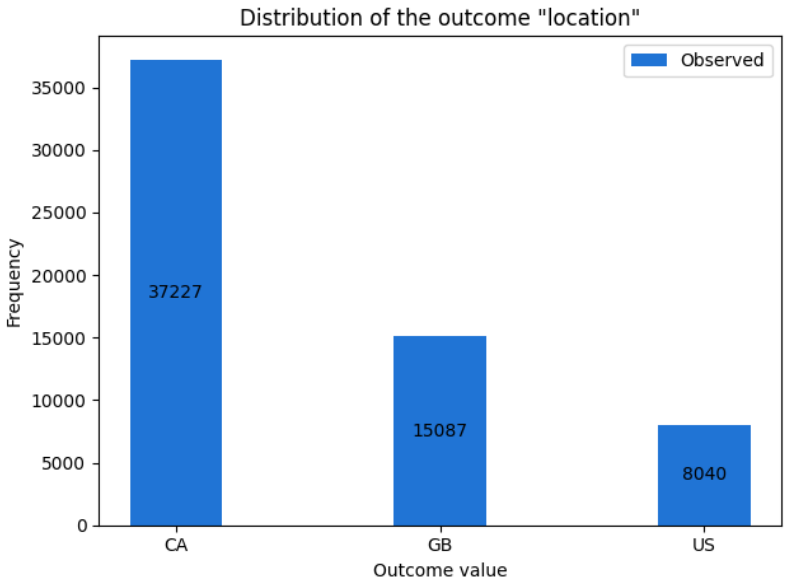
1. Analysis configuration
2. Pretraining bias metrics

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.

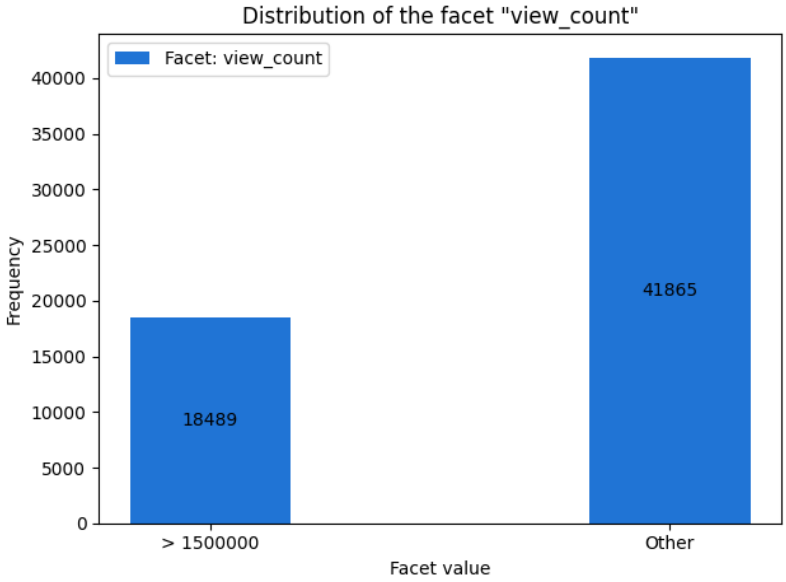
Outcome label: You chose the column `location` in the input data as the outcome label. Bias metric computation requires designating the positive outcome. You chose `location = CA` as the positive outcome. `location` consisted of values `['CA', 'GB', 'US']`.

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



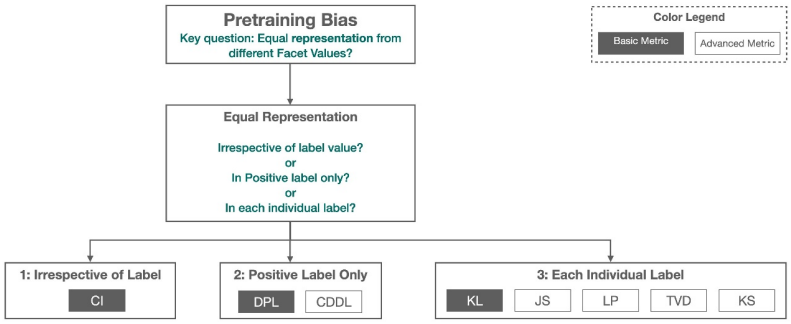
Facet: You chose the column `view_count` in the input data as the facet. `view_count` varied between `0.00` and `264407389.00`. Bias metrics were computed by comparing the inputs `view_count > 1500000` with all other inputs.

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



Pre-training Bias Metrics

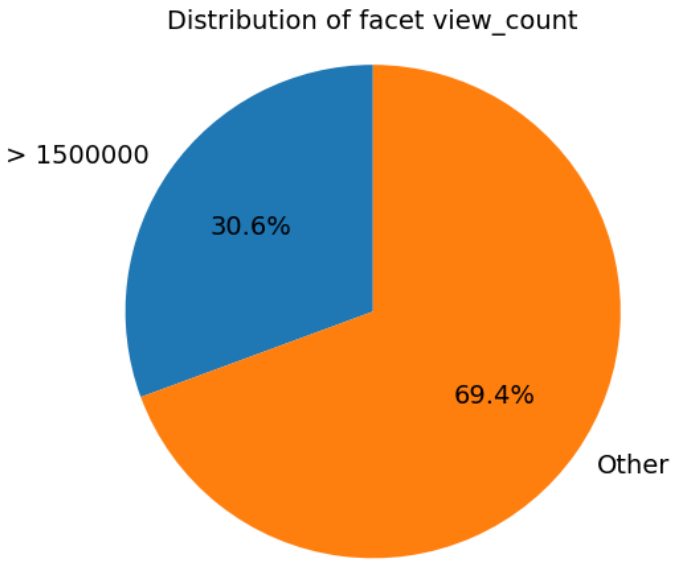
Pretraining bias metrics measure imbalances in facet value representation in the training data. Imbalances can be measured across different dimensions. For instance, you could focus imbalances within the inputs with positive observed label only. The figure below shows how different pretraining bias metrics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



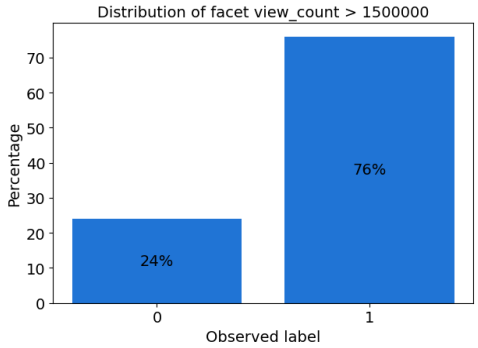
The metric values along with an informal description of what they mean are shown below. For mathematical formulas and examples, see the Measure Pretraining Bias section of the AWS documentation.

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

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



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



Facet Value(s)/Threshold: view_count > 1500000

Metric	Description	Value
Class Imbalance (CI)	Measures the imbalance in the number of inputs with facet values Sex=0 and rest of the inputs.	0.387
Difference in Proportions of Labels (DPL)	Measures the imbalance of positive observed labels between facet values Sex=0 and rest of the inputs.	-0.207
Jensen-Shannon Divergence (JS)	Measures how much the observed label distributions of facet values Sex=0 and rest of the inputs diverge from each other entropically.	0.024
Kullback-Leibler Divergence (KL)	Measures how much the observed label distributions of facet values Sex=0 and rest of the inputs diverge from each other entropically.	0.102
Kolmogorov-Smirnov (KS)	Measures maximum divergence between the observed label distributions for facet values Sex=0 and rest of the inputs in the dataset.	0.207
Lp-norm (LP)	Measures a p-norm difference between the observed label distributions associated with facet values Sex=0 rest of the inputs in the dataset.	0.292
Total Variation Distance (TVD)	Measures half of the L1-norm difference between the observed label distributions associated with facet values Sex=0 and rest of the inputs in the dataset.	0.207

Appendix: Analysis Configuration Parameters


```
{
   "dataset_type": "text/csv",
   "headers": [
      "location",
      "title",
      "channelTitle",
      "tags",
      "view_count",
      "likes",
```

```
        "likes",
        "dislikes",
        "comment_count",
        "comments_disabled",
        "ratings_disabled",
        "description_length",
        "timepublished",
        "trending_month",
        "trending_year",
        "category"
    ],
    "label": "location",
    "label_values_or_threshold": [
        "CA"
    ],
    "facet": [
        {
            "name_or_index": "view_count",
            "value_or_threshold": [
                1500000
            ]
        }
    ],
    "methods": {
        "report": {
            "name": "report",
            "title": "Analysis Report"
        },
        "pre_training_bias": {
            "methods": [
                "CI",
                "DPL",
                "KL",
                "JS",
                "LP",
                "TVD",
                "KS"
            ]
        }
    }
}
```