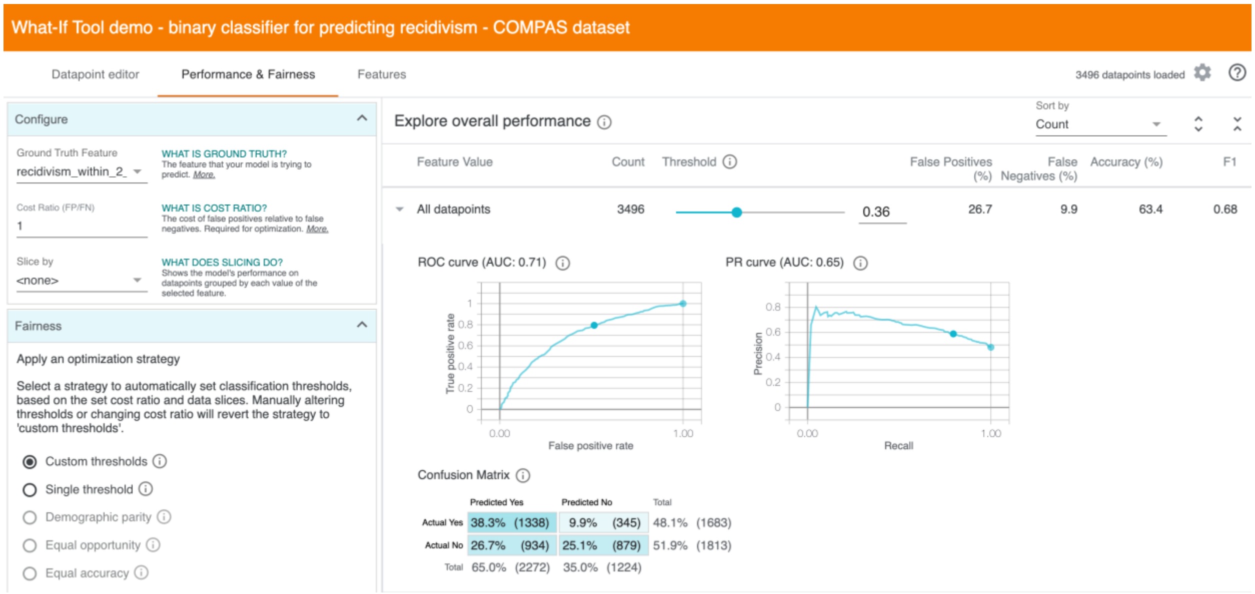
Assignment # 6:	What if Tool
Course:	CAP 6610 Applied Machine Learning
Professor:	Dr. Oge Marques
Student Number:	Z23596812
Student Name:	Renee Raven
Term:	Fall 2022

Report Part 1:

Demo 1: Investigate fairness on recidivism classification (Web demos)



Ground Truth Feature: recidivism_within_2_

WHAT IS GROUND TRUTH? The feature that your model is trying to predict. More.

Cost Ratio (FP/FN): 1

WHAT IS COST RATIO? The cost of false positives relative to false negatives. Required for optimization. More.

Slice by: <none>

WHAT DOES SLICING DO? Shows the model's performance on datapoints grouped by each value of the selected feature.

Apply an optimization strategy

Select a strategy to automatically set classification thresholds, based on the set cost ratio and data slices. Manually altering thresholds or changing cost ratio will revert the strategy to 'custom thresholds'.

● Custom thresholds
○ Single threshold
○ Demographic parity
○ Equal opportunity
○ Equal accuracy

Sort by: Count

Feature Value	Count	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
All datapoints	3496	0.36	26.7	9.9	63.4	0.68

ROC curve (AUC: 0.71)

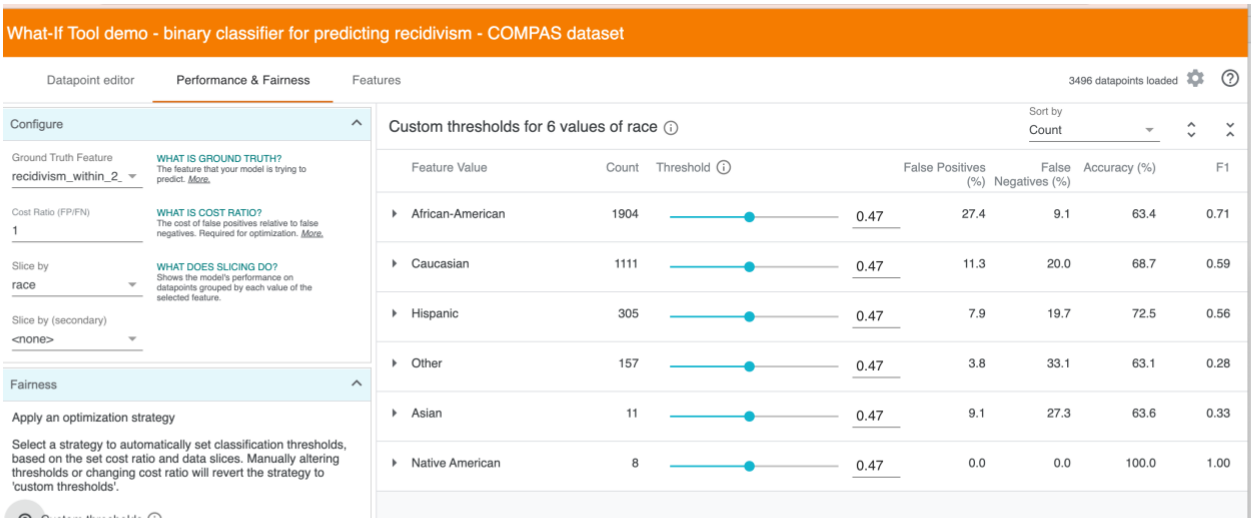
PR curve (AUC: 0.65)

Confusion Matrix

	Predicted Yes	Predicted No	Total
Actual Yes	38.3% (1338)	9.9% (345)	48.1% (1683)
Actual No	26.7% (934)	25.1% (879)	51.9% (1813)
Total	65.0% (2272)	35.0% (1224)	

This binary classification uses the popular COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm for predicting recidivism. ProPublica has demonstrated that this algorithm propagates racial bias, favoring white offenders and disproportionately predicting recidivism for black offenders. The dataset contains 3496 datapoints from the 10,000 observations available in the dataset of criminal defendants in Broward County, Florida. It is unclear how this sample was culled from the larger pool. The Ground truth is recidivism within 2 years.

Looking at a slice by race model with a standard Cost Ratio of 1 reveals a disturbing and disproportionate number of False Positives for those of African-American race (27.4%) versus those of Caucasian race (11.3%). This indicates those offenders were falsely predicted to reoffend within 2 years, but they did not. It is also notable that the False Negative rate was over double for those of Caucasian race (20.0%) versus those of African-American race (9.1%). This confirms that the algorithm's was more than twice as likely to misidentify those of Caucasian race as low risk compared to those of African-American race.



What-If Tool demo - binary classifier for predicting recidivism - COMPAS dataset

Configure

Ground Truth Feature: recidivism_within_2_

WHAT IS GROUND TRUTH? The feature that your model is trying to predict. *More.*

Cost Ratio (FP/FN): 1

WHAT IS COST RATIO? The cost of false positives relative to false negatives. Required for optimization. *More.*

Slice by: race

WHAT DOES SLICING DO? Shows the model's performance on datapoints grouped by each value of the selected feature.

Slice by (secondary): <none>

Fairness

Apply an optimization strategy

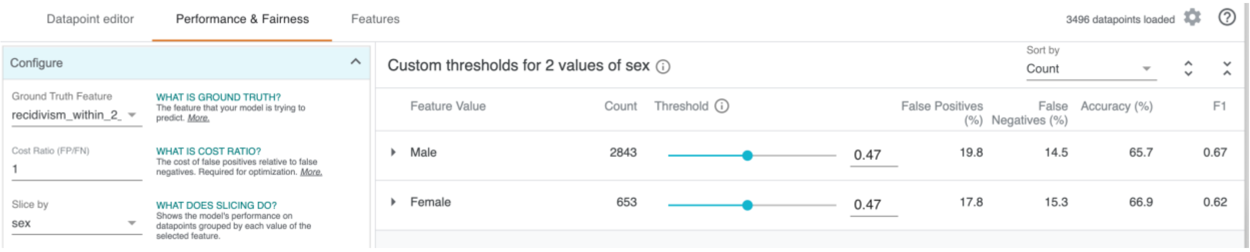
Select a strategy to automatically set classification thresholds, based on the set cost ratio and data slices. Manually altering thresholds or changing cost ratio will revert the strategy to 'custom thresholds'.

Custom thresholds for 6 values of race

Sort by: Count

Feature Value	Count	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
African-American	1904	0.47	27.4	9.1	63.4	0.71
Caucasian	1111	0.47	11.3	20.0	68.7	0.59
Hispanic	305	0.47	7.9	19.7	72.5	0.56
Other	157	0.47	3.8	33.1	63.1	0.28
Asian	11	0.47	9.1	27.3	63.6	0.33
Native American	8	0.47	0.0	0.0	100.0	1.00

When the same data is sliced by gender, we note the False Positives and False Negatives are quite similar (Male 19.8% and Female 17.8%).



Configure

Ground Truth Feature: recidivism_within_2_

WHAT IS GROUND TRUTH? The feature that your model is trying to predict. *More.*

Cost Ratio (FP/FN): 1

WHAT IS COST RATIO? The cost of false positives relative to false negatives. Required for optimization. *More.*

Slice by: sex

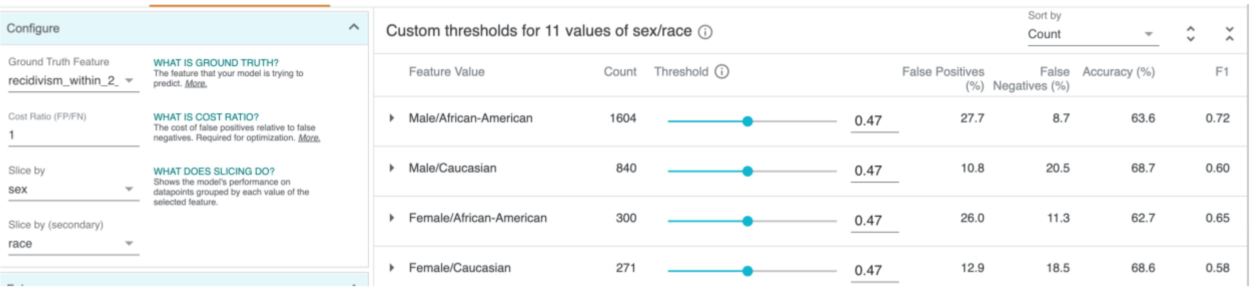
WHAT DOES SLICING DO? Shows the model's performance on datapoints grouped by each value of the selected feature.

Custom thresholds for 2 values of sex

Sort by: Count

Feature Value	Count	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
Male	2843	0.47	19.8	14.5	65.7	0.67
Female	653	0.47	17.8	15.3	66.9	0.62

When the data is also sliced by race (as a secondary slice), we discover the bias occurs for both Male and Female African-Americans. Once again, the False Positive rate is much higher for those in the African-American group and the False Negative higher for those in the Caucasian group.



Configure

Ground Truth Feature: recidivism_within_2_

WHAT IS GROUND TRUTH? The feature that your model is trying to predict. *More.*

Cost Ratio (FP/FN): 1

WHAT IS COST RATIO? The cost of false positives relative to false negatives. Required for optimization. *More.*

Slice by: sex

WHAT DOES SLICING DO? Shows the model's performance on datapoints grouped by each value of the selected feature.

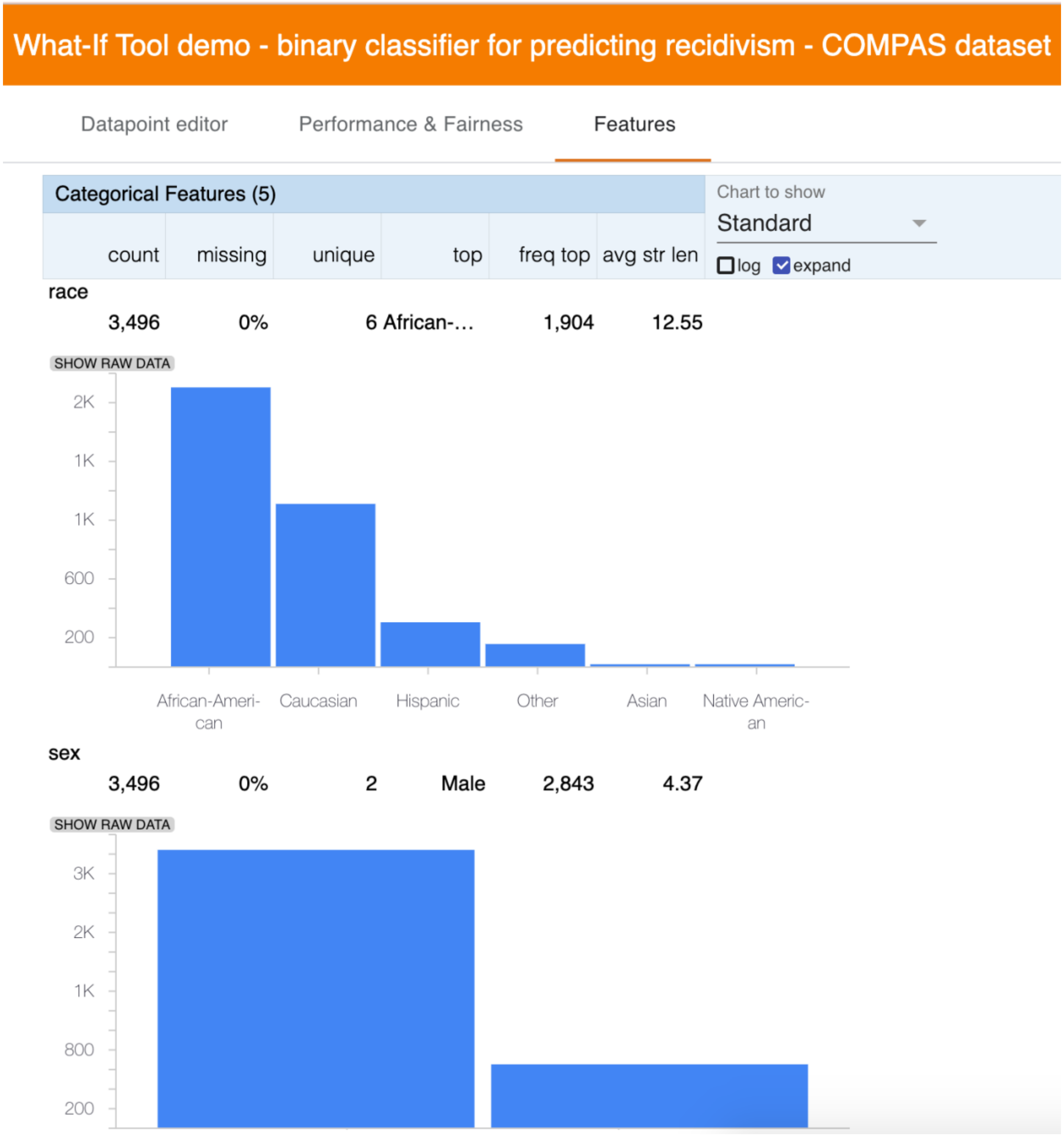
Slice by (secondary): race

Custom thresholds for 11 values of sex/race

Sort by: Count

Feature Value	Count	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
Male/African-American	1604	0.47	27.7	8.7	63.6	0.72
Male/Caucasian	840	0.47	10.8	20.5	68.7	0.60
Female/African-American	300	0.47	26.0	11.3	62.7	0.65
Female/Caucasian	271	0.47	12.9	18.5	68.6	0.58

The dataset is unbalanced for both race and sex, but there are adequate numbers of samples to compare Caucasian and African-Americans (1111 examples of those of Caucasian race and 1904 examples of those of African-American race).



What-If Tool demo - binary classifier for predicting recidivism - COMPAS dataset

Chart to show: Standard

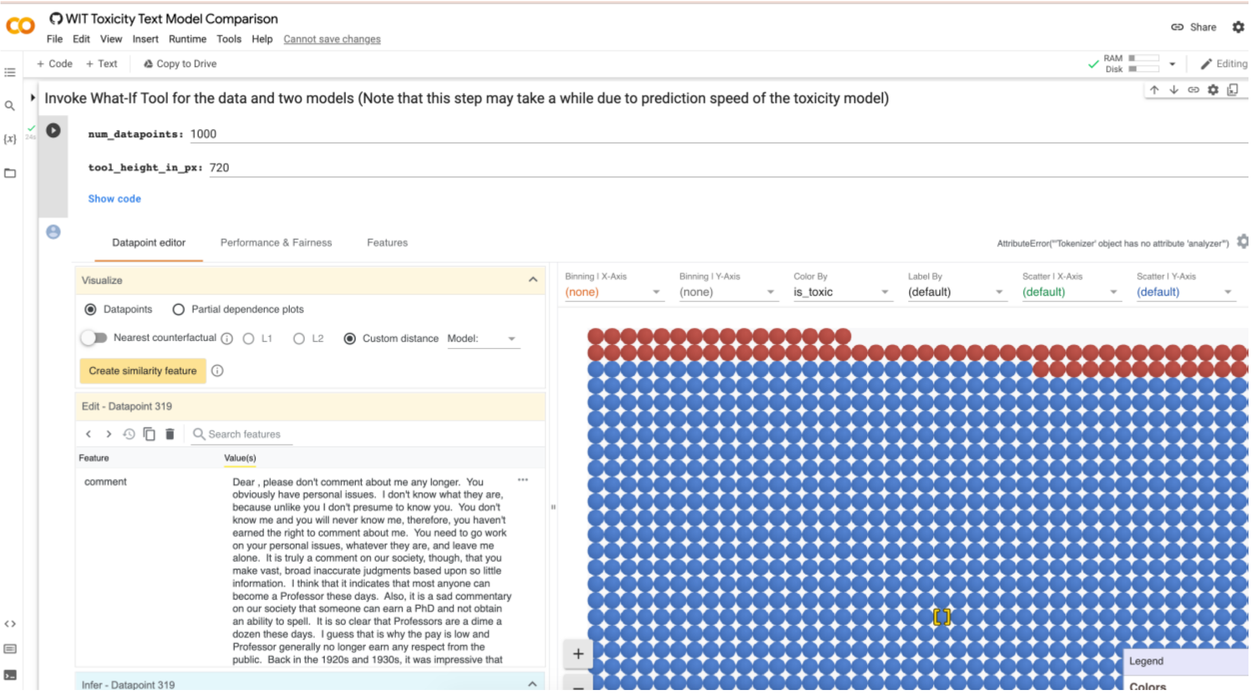
□ log ☑ expand

Categorical Features (5)

	count	missing	unique	top	freq top	avg str len
race	3,496	0%	6	African-…	1,904	12.55
sex	3,496	0%	2	Male	2,843	4.37

While this is an interesting example, I am troubled that I could not find an explanation for how the 10,000 samples were whittled down to 3496 observations. It suggests that the selection of instances could be partly to blame for the bias.

Demo 2: Text toxicity classifiers (Notebook demos)


WIT Toxicity Text Model Comparison
File Edit View Insert Runtime Tools Help Cannot save changes
Share
+ Code + Text Copy to Drive
RAM
Disk
Editing
Show code
Datapoint editor
Performance & Fairness
Features
Visualize
Create similarity feature
Search features
Legend
Colors

Invoke What-If Tool for the data and two models (Note that this step may take a while due to prediction speed of the toxicity model)

num_datapoints: 1000

tool_height_in_px: 720

AttributeError("'Tokenizer' object has no attribute 'analyzer'")

● Datapoints ○ Partial dependence plots

Nearest counterfactual ○ L1 ○ L2 ● Custom distance Model:

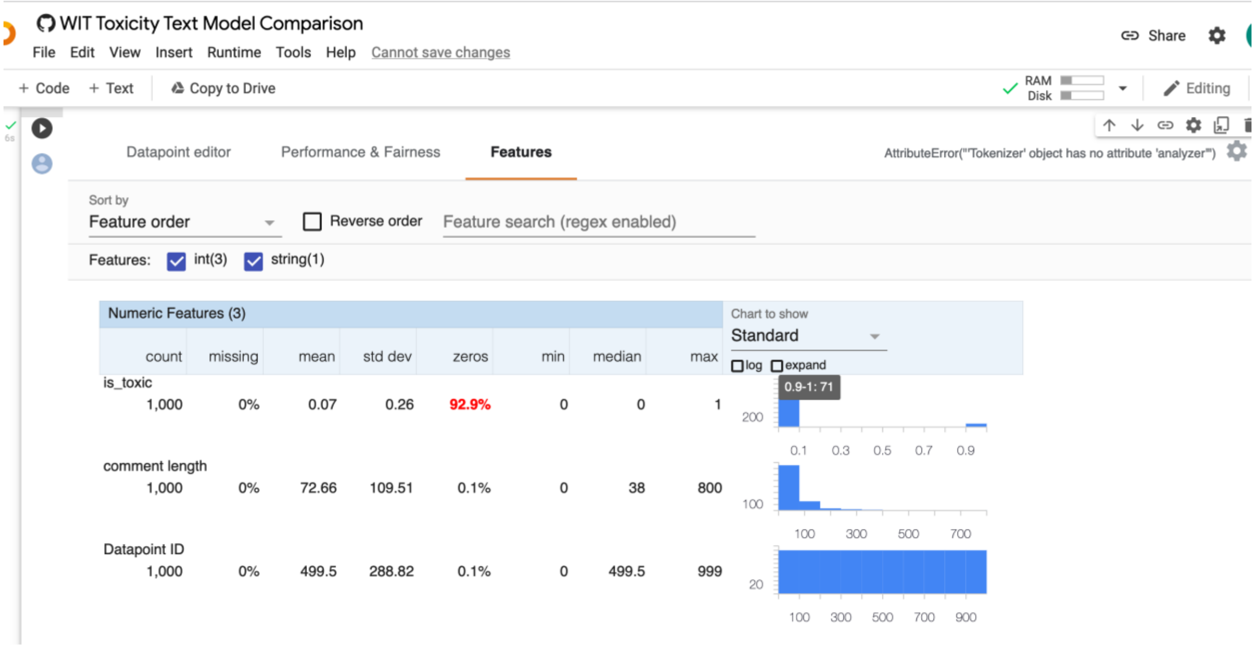
Binning | X-Axis (none)
Binning | Y-Axis (none)
Color By is_toxic
Label By (default)
Scatter | X-Axis (default)
Scatter | Y-Axis (default)

Edit - Datapoint 319

Feature	Value(s)
comment	Dear , please don't comment about me any longer. You obviously have personal issues. I don't know what they are, because unlike you I don't presume to know you. You don't know me and you will never know me, therefore, you haven't earned the right to comment about me. You need to go work on your personal issues, whatever they are, and leave me alone. It is truly a comment on our society, though, that you make vast, broad inaccurate judgments based upon so little information. I think that it indicates that most anyone can become a Professor these days. Also, it is a sad commentary on our society that someone can earn a PhD and not obtain an ability to spell. It is so clear that Professors are a dime a dozen these days. I guess that is why the pay is low and Professor generally no longer earn any respect from the public. Back in the 1920s and 1930s, it was impressive that

Infer - Datapoint 319

The text toxicity classifiers model appeared to be an interesting model using ConversationAI to determine text toxicity. There are 1000 unique observations, 929 labelled 0 for is_toxic (meaning not toxic) and 71 labelled 1 for is_toxic.



AttributeError("'Tokenizer' object has no attribute 'analyzer'")

Sort by Feature order ☐ Reverse order Feature search (regex enabled)

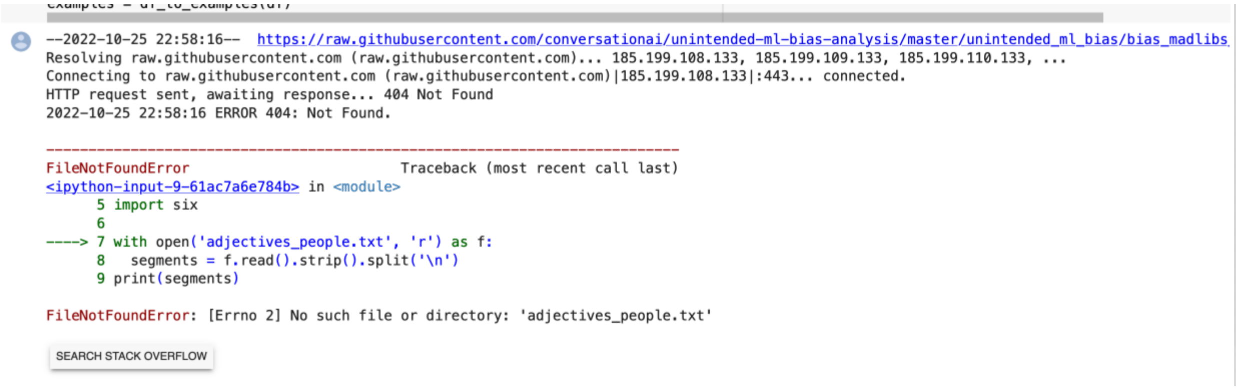
Features: ☑ int(3) ☑ string(1)

Chart to show Standard ☐log ☐expand

Numeric Features (3)

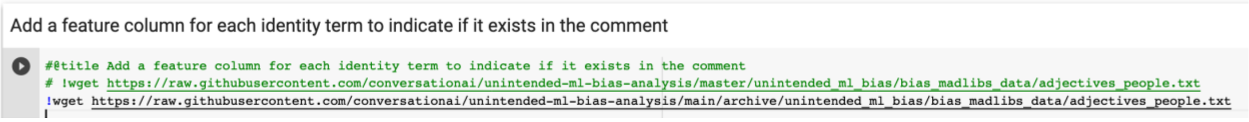
	count	missing	mean	std dev	zeros	min	median	max
is_toxic	1,000	0%	0.07	0.26	92.9%	0	0	1
comment length	1,000	0%	72.66	109.51	0.1%	0	38	800
Datapoint ID	1,000	0%	499.5	288.82	0.1%	0	499.5	999

However, running this model proved an insurmountable challenge for me. The first error involved a mistaken path for the adjectives_people.txt file.



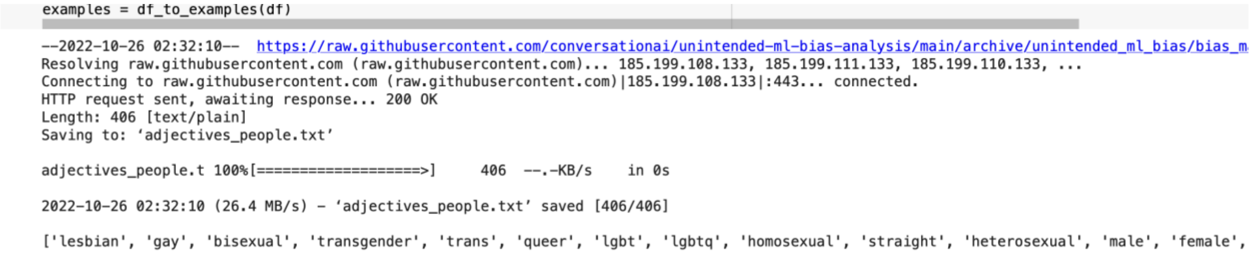
```
examples = df_to_examples(df)

--2022-10-25 22:58:16--  https://raw.githubusercontent.com/conversationai/unintended-ml-bias-analysis/master/unintended_ml_bias/bias_madlibs
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.133, 185.199.110.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... connected.
HTTP request sent, awaiting response... 404 Not Found
2022-10-25 22:58:16 ERROR 404: Not Found.

---------------------------------------------------------------------------
FileNotFoundError                         Traceback (most recent call last)
<ipython-input-9-61ac7a6e784b> in <module>
      5 import six
      6 
----> 7 with open('adjectives_people.txt', 'r') as f:
      8   segments = f.read().strip().split('\n')
      9 print(segments)

FileNotFoundError: [Errno 2] No such file or directory: 'adjectives_people.txt'
```

After finding the correct path for the file and replacing it in the code, that part was able to run.



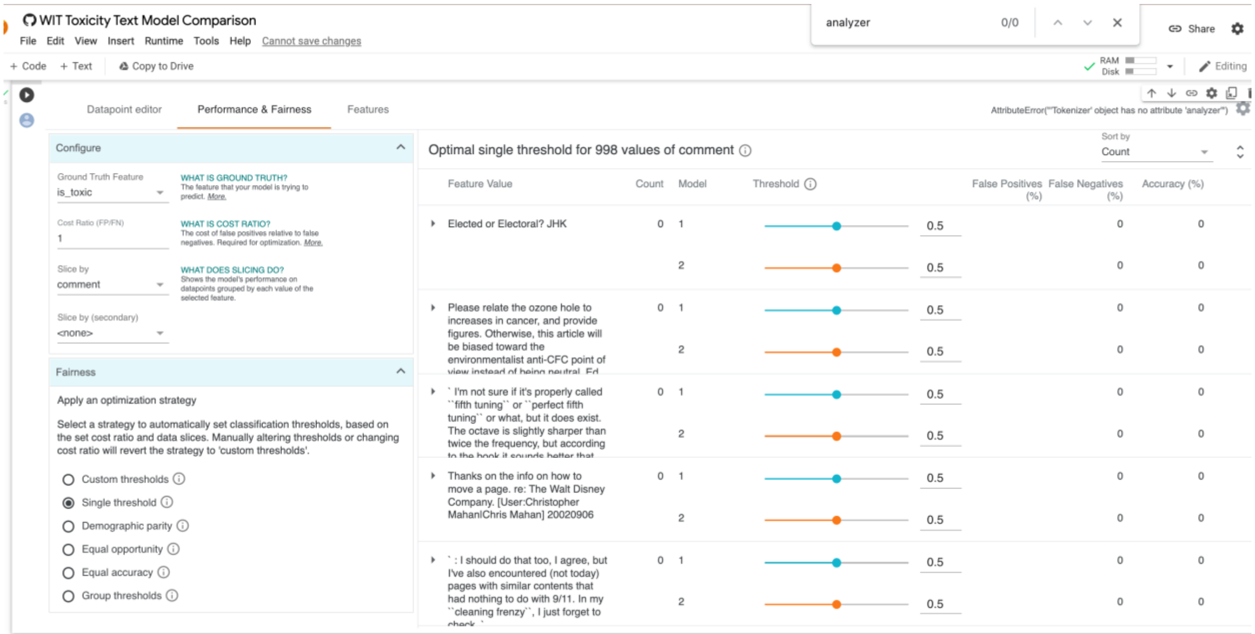
Add a feature column for each identity term to indicate if it exists in the comment

```
#@title Add a feature column for each identity term to indicate if it exists in the comment
# !wget https://raw.githubusercontent.com/conversationai/unintended-ml-bias-analysis/master/unintended_ml_bias/bias_madlibs_data/adjectives_people.txt
!wget https://raw.githubusercontent.com/conversationai/unintended-ml-bias-analysis/main/archive/unintended_ml_bias/bias_madlibs_data/adjectives_people.txt
```



```
examples = df_to_examples(df)

--2022-10-26 02:32:10--  https://raw.githubusercontent.com/conversationai/unintended-ml-bias-analysis/main/archive/unintended_ml_bias/bias_m
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.111.133, 185.199.110.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 406 [text/plain]
Saving to: ‘adjectives_people.txt’

adjectives_people.t 100%[===================>]     406  --.-KB/s    in 0s

2022-10-26 02:32:10 (26.4 MB/s) - ‘adjectives_people.txt’ saved [406/406]

['lesbian', 'gay', 'bisexual', 'transgender', 'trans', 'queer', 'lgbt', 'lgbtq', 'homosexual', 'straight', 'heterosexual', 'male', 'female',
```

Next I set the Ground Truth to is_toxic, but I was not able to generate results, even when attempting to slice the data or change the threshold.


WIT Toxicity Text Model Comparison
File Edit View Insert Runtime Tools Help Cannot save changes
+ Code + Text Copy to Drive
analyzer 0/0
Share
RAM Disk
Editing
Datapoint editor
Performance & Fairness
Features

AttributeError("'Tokenizer' object has no attribute 'analyzer'")

Configure

Ground Truth Feature: is_toxic

WHAT IS GROUND TRUTH? The feature that your model is trying to predict. More.

Cost Ratio (FP/FN): 1

WHAT IS COST RATIO? The cost of false positives relative to false negatives. Required for optimization. More.

Slice by: comment

WHAT DOES SLICING DO? Shows the model's performance on datapoints grouped by each value of the selected feature.

Slice by (secondary): <none>

Fairness

Apply an optimization strategy

Select a strategy to automatically set classification thresholds, based on the set cost ratio and data slices. Manually altering thresholds or changing cost ratio will revert the strategy to 'custom thresholds'.

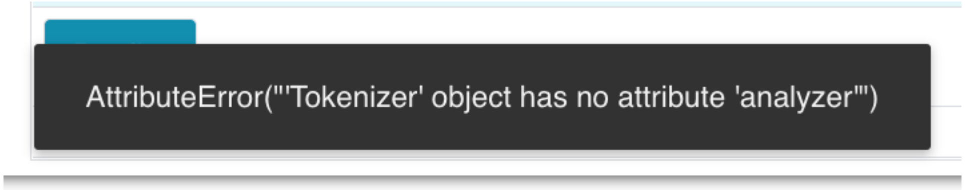
○ Custom thresholds
● Single threshold
○ Demographic parity
○ Equal opportunity
○ Equal accuracy
○ Group thresholds

Optimal single threshold for 998 values of comment

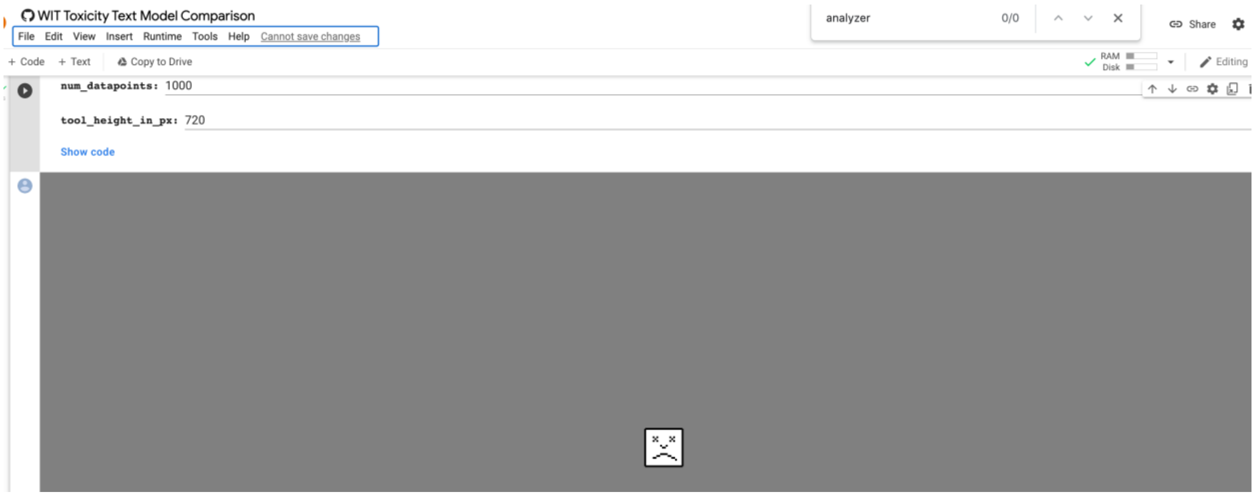
Sort by: Count

Feature Value	Count	Model	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)
Elected or Electoral? JHK	0	1	0.5		0	0
		2	0.5		0	0
Please relate the ozone hole to increases in cancer, and provide figures. Otherwise, this article will be biased toward the environmentalist anti-CFC point of view instead of being neutral. Ed	0	1	0.5		0	0
		2	0.5		0	0
` I'm not sure if it's properly called ``fifth tuning`` or ``perfect fifth tuning`` or what, but it does exist. The octave is slightly sharper than twice the frequency, but according to the book it sounds better that	0	1	0.5		0	0
		2	0.5		0	0
Thanks on the info on how to move a page. re: The Walt Disney Company. [User:Christopher MahanlChris Mahan] 20020906	0	1	0.5		0	0
		2	0.5		0	0
` : I should do that too, I agree, but I've also encountered (not today) pages with similar contents that had nothing to do with 9/11. In my ``cleaning frenzy``, I just forget to check. `	0	1	0.5		0	0
		2	0.5		0	0

The main problem appeared to involve the tokenizer and I wasn't able to find a solution online.



AttributeError("'Tokenizer' object has no attribute 'analyzer'")


WIT Toxicity Text Model Comparison
File Edit View Insert Runtime Tools Help Cannot save changes
+ Code + Text Copy to Drive
analyzer 0/0
Share
RAM Disk
Editing
Show code

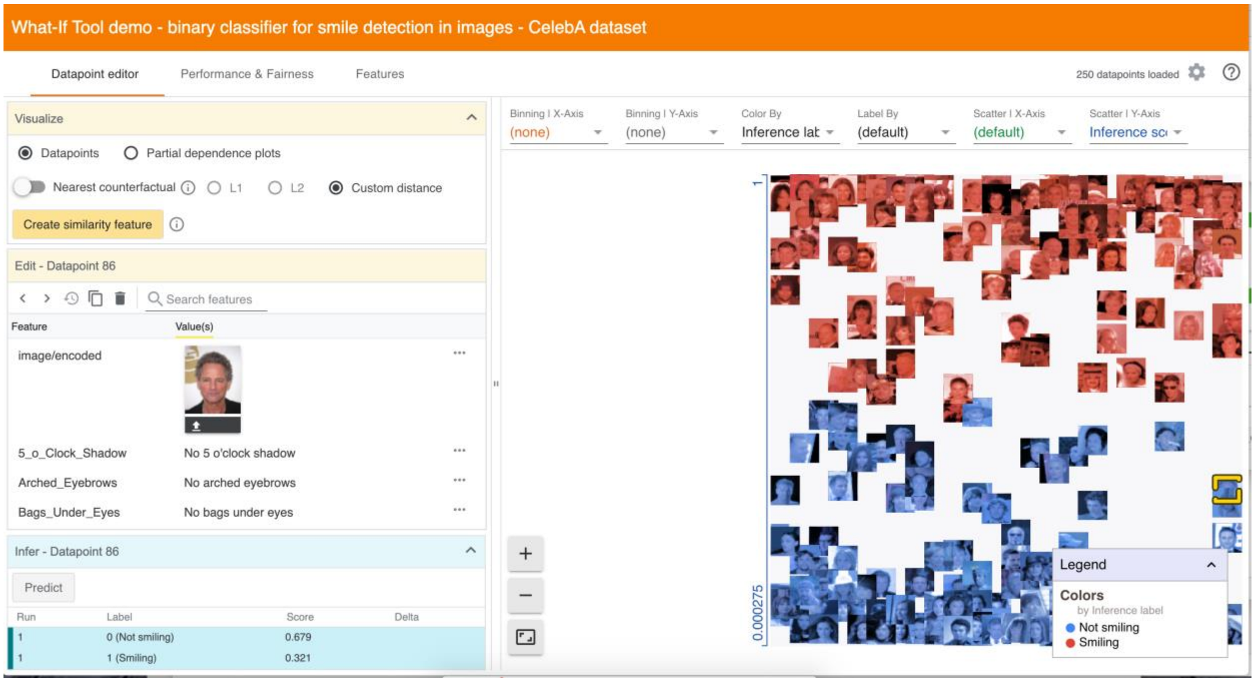
num_datapoints: 1000

tool_height_in_px: 720

Sometimes it is best to cut your losses and move on to the next opportunity.

Demo 3: Binary classifier for smile detection in images – CelebA dataset (Web demos)

This model uses 250 images from the CelebA binary dataset (the original dataset has over 200,000 images). Thirty of the original 40 binary features/attributes are included. There are 9 missing binary features/attributes and 5 missing landmarks from the original dataset. The description offers a fun challenge: “Can you identify which group was missing from the training data, resulting in a biased model?” This also informs us we are using a biased dataset.



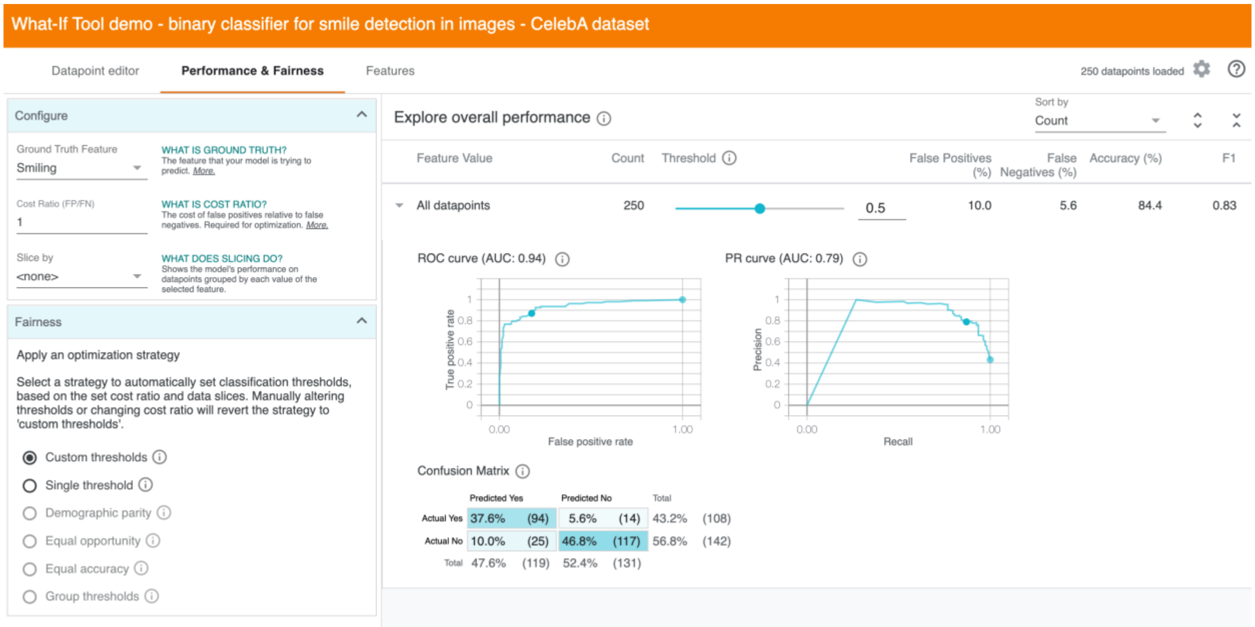
What-If Tool demo - binary classifier for smile detection in images - CelebA dataset

Binning | X-Axis (none)
Binning | Y-Axis (none)
Color By Inference lat
Label By (default)
Scatter | X-Axis (default)
Scatter | Y-Axis Inference sc

Feature	Value(s)
image/encoded	
5_o_Clock_Shadow	No 5 o'clock shadow
Arched_Eyebrows	No arched eyebrows
Bags_Under_Eyes	No bags under eyes

Run	Label	Score	Delta
1	0 (Not smiling)	0.679	
1	1 (Smiling)	0.321	

The Datapoint editor offers a collection of photos/datapoints to select and provides a prediction for the image selected. Here we see that it correctly predicted Lindsey Buckingham is Not Smiling.



What-If Tool demo - binary classifier for smile detection in images - CelebA dataset

Datapoint editor | **Performance & Fairness** | Features

250 datapoints loaded

Configure

Ground Truth Feature: Smiling

WHAT IS GROUND TRUTH?
The feature that your model is trying to predict. *More.*

Cost Ratio (FP/FN): 1

WHAT IS COST RATIO?
The cost of false positives relative to false negatives. Required for optimization. *More.*

Slice by: <none>

WHAT DOES SLICING DO?
Shows the model's performance on datapoints grouped by each value of the selected feature.

Fairness

Apply an optimization strategy

Select a strategy to automatically set classification thresholds, based on the set cost ratio and data slices. Manually altering thresholds or changing cost ratio will revert the strategy to 'custom thresholds'.

- ● Custom thresholds
- ○ Single threshold
- ○ Demographic parity
- ○ Equal opportunity
- ○ Equal accuracy
- ○ Group thresholds

Explore overall performance

Sort by: Count

Feature Value	Count	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
All datapoints	250	0.5	10.0	5.6	84.4	0.83

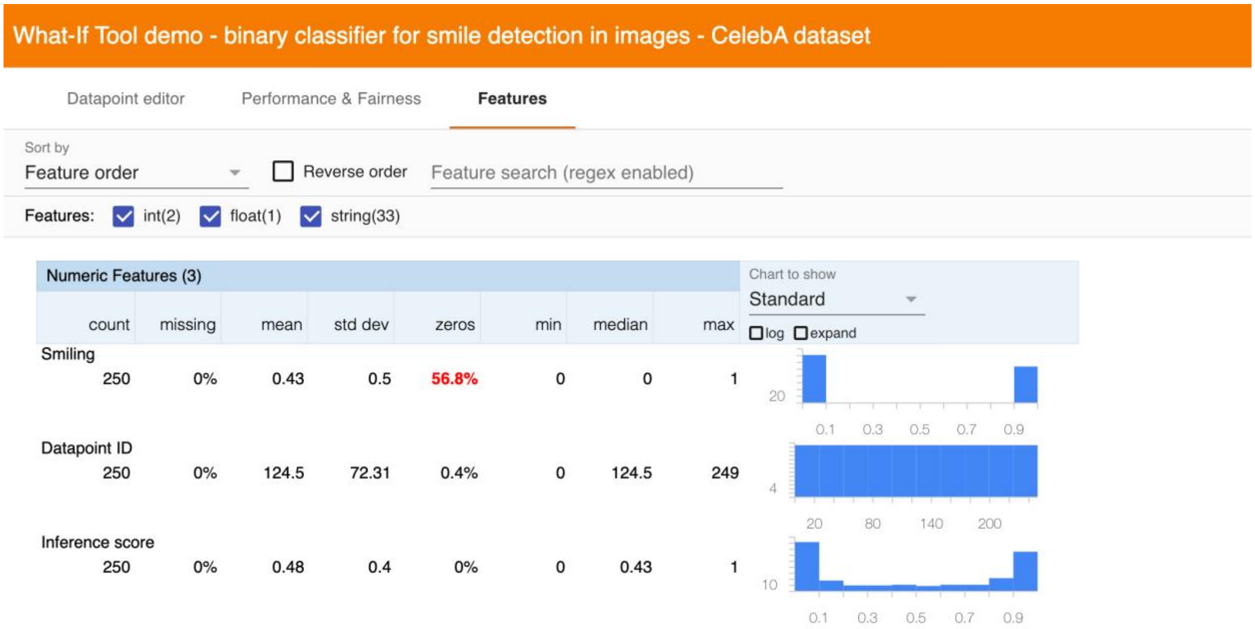
ROC curve (AUC: 0.94)

PR curve (AUC: 0.79)

Confusion Matrix

	Predicted Yes		Predicted No		Total	
Actual Yes	37.6%	(94)	5.6%	(14)	43.2%	(108)
Actual No	10.0%	(25)	46.8%	(117)	56.8%	(142)
Total	47.6%	(119)	52.4%	(131)		

The model using Smiling as the ground truth performs fairly well, demonstrating an overall 84.4% accuracy and 0.94 ROC curve.



What-If Tool demo - binary classifier for smile detection in images - CelebA dataset

Datapoint editor | Performance & Fairness | **Features**

Sort by: Feature order ☐ Reverse order Feature search (regex enabled)

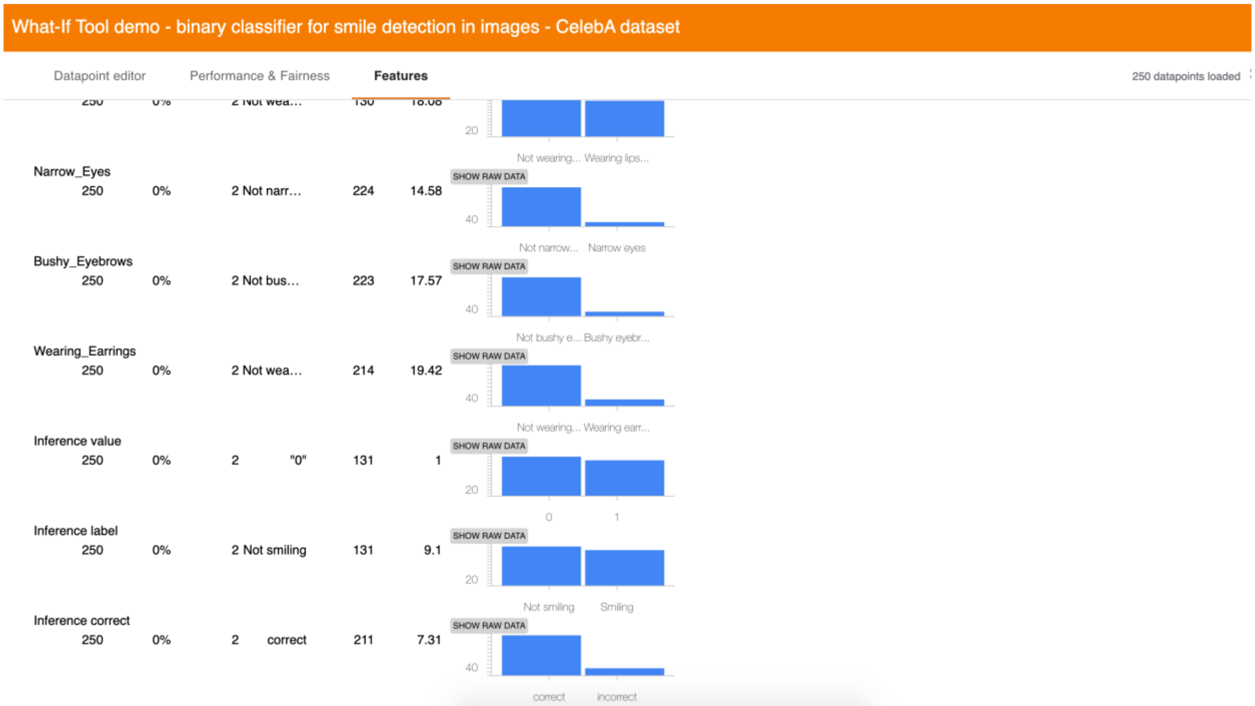
Features: ☑ int(2) ☑ float(1) ☑ string(33)

Numeric Features (3)

Chart to show: Standard ☐ log ☐ expand

	count	missing	mean	std dev	zeros	min	median	max
Smiling	250	0%	0.43	0.5	56.8%	0	0	1
Datapoint ID	250	0%	124.5	72.31	0.4%	0	124.5	249
Inference score	250	0%	0.48	0.4	0%	0	0.43	1

Additionally, the dataset is split around 56% - 44% for the Not Smiling and Smiling classes. The graph for the inference score demonstrates an interesting reverse bell curve like shape.



What-If Tool demo - binary classifier for smile detection in images - CelebA dataset

	250	0%	2 Not wea…	130	18.08
Narrow_Eyes					
	250	0%	2 Not narr…	224	14.58
Bushy_Eyebrows					
	250	0%	2 Not bus…	223	17.57
Wearing_Earrings					
	250	0%	2 Not wea…	214	19.42
Inference value					
	250	0%	2 "0"	131	1
Inference label					
	250	0%	2 Not smiling	131	9.1
Inference correct					
	250	0%	2 correct	211	7.31

The features were not so evenly split. Some were near 50% for each, but most included more observations in one class over another.

Report Part 2:

I selected the smile detection web based demo because it offered a fun puzzle (“Can you identify which group was missing from the training data, resulting in a biased model?”) on an admittedly biased dataset. From the list of features there are 39 features plus the Smiling target in the original dataset:

```
5_o_Clock_Shadow Arched_Eyebrows Attractive Bags_Under_Eyes Bald Bangs
Big_Lips Big_Nose Black_Hair Blond_Hair Blurry Brown_Hair Bushy_Eyebrows
Chubby Double_Chin Eyeglasses Goatee Gray_Hair Heavy_Makeup
High_Cheekbones Male Mouth_Slightly_Open Mustache Narrow_Eyes No_Beard
Oval_Face Pale_Skin Pointy_Nose Receding_Hairline Rosy_Cheeks Sideburns
Smiling Straight_Hair Wavy_Hair Wearing_Earrings Wearing_Hat
Wearing_Lipstick Wearing_Necklace Wearing_Necktie Young
```

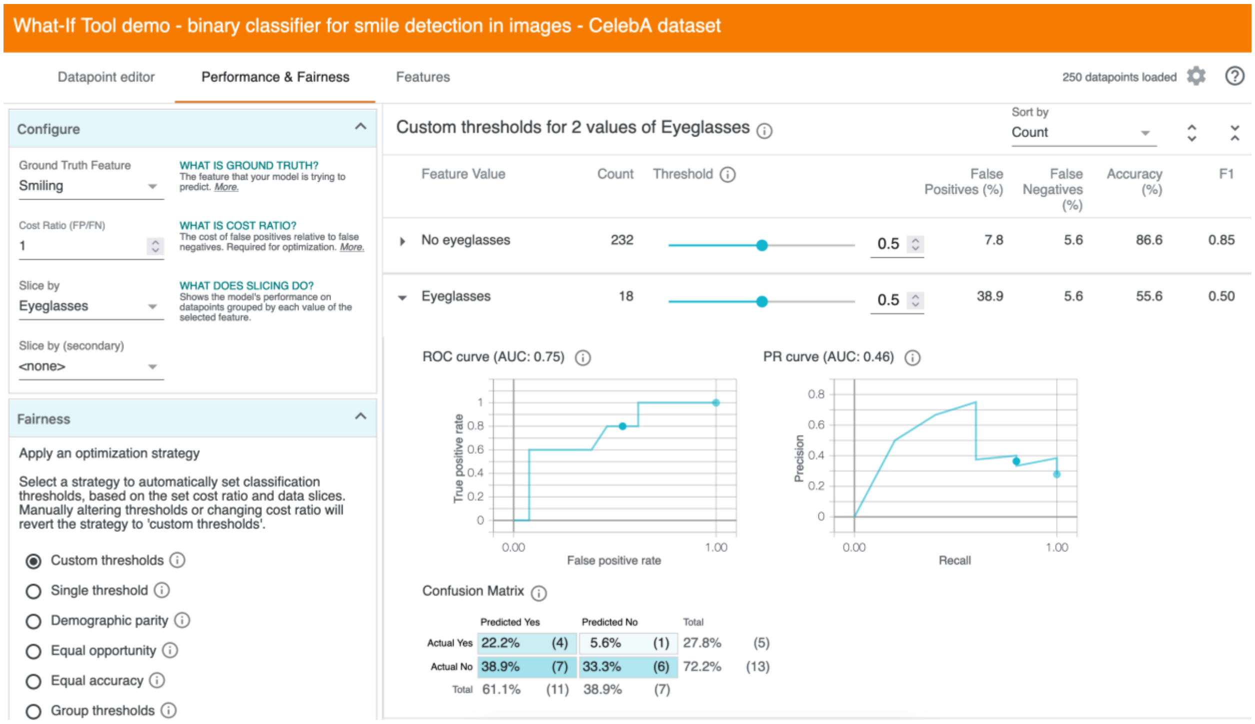
Performing a manual comparison, the missing features in our demo are:

Attractive
Big Lips
Big Nose
Chubby
Double Chin
Heavy Makeup
Pointy Nose

Receding Hairline
Rosy Cheeks

Given human nature, each of these 9 features could sway people to perceive the facial characteristics differently. People deemed attractive or rosy cheeked may be more likely to illicit a favorable response from viewers and the False Positive rate might be higher (perceiving smiling when they are not). Similarly, characteristic often associated with people deemed less attractive (chubby, pointy nose, heavy makeup) may increase the number of False Negatives where the model predicts no smile even when there is one. There may also be a racial factor as big lips are sometimes associated with individuals with African-American heritage.

It must be noted that all of this is based on the model being trained on human labelled data. Then these human biases are manifested in a biased dataset.



What-If Tool demo - binary classifier for smile detection in images - CelebA dataset

Configure

Ground Truth Feature: Smiling

WHAT IS GROUND TRUTH? The feature that your model is trying to predict. More.

Cost Ratio (FP/FN): 1

WHAT IS COST RATIO? The cost of false positives relative to false negatives. Required for optimization. More.

Slice by: Eyeglasses

WHAT DOES SLICING DO? Shows the model's performance on datapoints grouped by each value of the selected feature.

Slice by (secondary): <none>

Fairness

Apply an optimization strategy

Select a strategy to automatically set classification thresholds, based on the set cost ratio and data slices. Manually altering thresholds or changing cost ratio will revert the strategy to 'custom thresholds'.

● Custom thresholds
○ Single threshold
○ Demographic parity
○ Equal opportunity
○ Equal accuracy
○ Group thresholds

Custom thresholds for 2 values of Eyeglasses

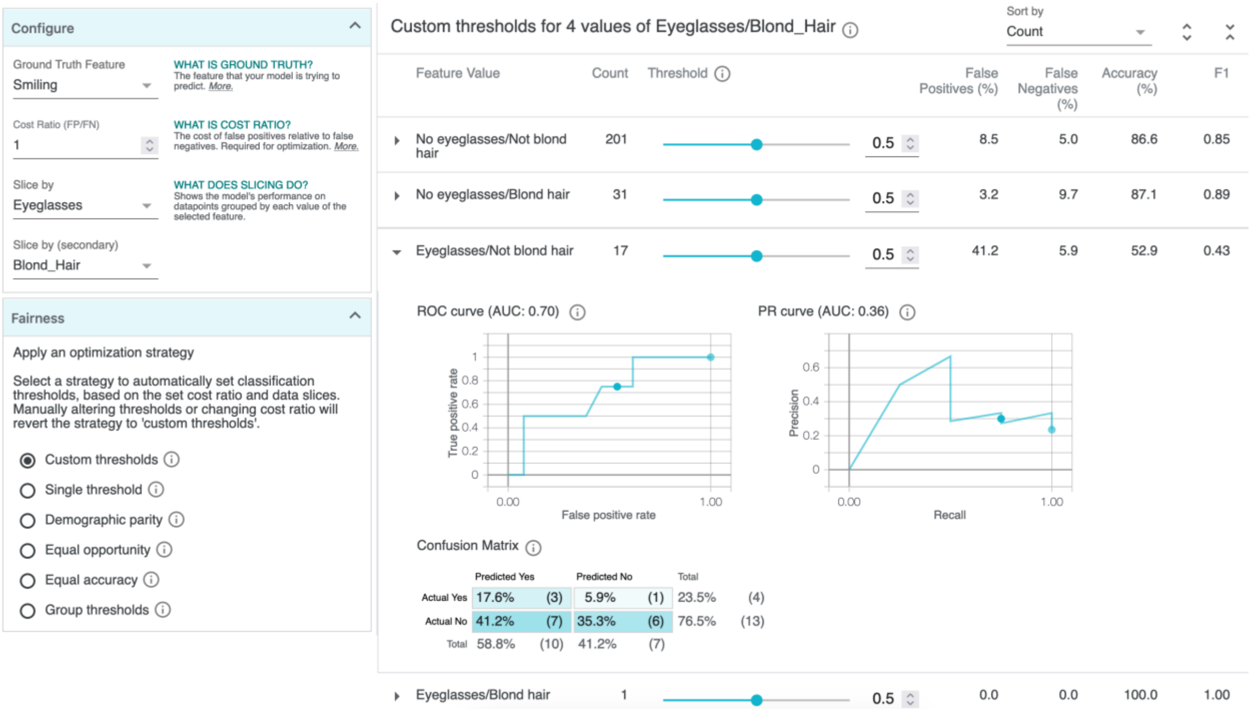
Sort by: Count

Feature Value	Count	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
No eyeglasses	232	0.5	7.8	5.6	86.6	0.85
Eyeglasses	18	0.5	38.9	5.6	55.6	0.50

Confusion Matrix

	Predicted Yes		Predicted No		Total	
Actual Yes	22.2%	(4)	5.6%	(1)	27.8%	(5)
Actual No	38.9%	(7)	33.3%	(6)	72.2%	(13)
Total	61.1%	(11)	38.9%	(7)		

The model appears quite confused by eyeglasses. The False Positive rate is close to 40% when the data is sliced on eyeglasses.



Configure

Ground Truth Feature: Smiling

WHAT IS GROUND TRUTH? The feature that your model is trying to predict. *More.*

Cost Ratio (FP/FN): 1

WHAT IS COST RATIO? The cost of false positives relative to false negatives. Required for optimization. *More.*

Slice by: Eyeglasses

WHAT DOES SLICING DO? Shows the model's performance on datapoints grouped by each value of the selected feature.

Slice by (secondary): Blond_Hair

Fairness

Apply an optimization strategy

Select a strategy to automatically set classification thresholds, based on the set cost ratio and data slices. Manually altering thresholds or changing cost ratio will revert the strategy to 'custom thresholds'.

- ● Custom thresholds
- ○ Single threshold
- ○ Demographic parity
- ○ Equal opportunity
- ○ Equal accuracy
- ○ Group thresholds

Custom thresholds for 4 values of Eyeglasses/Blond_Hair

Sort by: Count

Feature Value	Count	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
No eyeglasses/Not blond hair	201	0.5	8.5	5.0	86.6	0.85
No eyeglasses/Blond hair	31	0.5	3.2	9.7	87.1	0.89
Eyeglasses/Not blond hair	17	0.5	41.2	5.9	52.9	0.43
Eyeglasses/Blond hair	1	0.5	0.0	0.0	100.0	1.00

Confusion Matrix

	Predicted Yes		Predicted No		Total	
Actual Yes	17.6%	(3)	5.9%	(1)	23.5%	(4)
Actual No	41.2%	(7)	35.3%	(6)	76.5%	(13)
Total	58.8%	(10)	41.2%	(7)		

The False Positive error worsens to over 41% when combining blond hair and eyeglasses.

While it could be fun to play more with this tool, overall it seems like the WIT is still a WIP. I'm not sure how much I would learn from spending more time looking at each component and adjusting the different settings.

I would be curious to know more about the recidivism algorithms. Supposedly, the COMPAS algorithm was/is widely used. I wonder if that is the only one in use, how other algorithms compare, and how much the judicial system depends on these to determine sentencing. This small example certainly suggests that AI reflects human biases already present in a system and should be properly vetted before use.

I also found the discussion of AI Fairness interesting. The brief articles offers 5 views of fairness to illustrate different people may have different ideas of what is fair. The Group unaware, for instance, believes disregarding the feature that introduces bias (such as gender) will obviate the bias. The Equal accuracy camp argues the model should be tuned to the feature so that the percentage of misclassifications will be the same for the divisive classes of the feature. In a clever turn of phrase, the article asks "Which sense of fairness is the fairest in the land?" This is a question worth much further inquiry.