Discriminatory Algorithms and Biased Data

Is the Future of Machine Learning Doomed?

Celeste Tretto, Data Scientist

Sarah Moir, Program Manager
October 4, 2018

Forward-Looking Statements

During the course of this presentation, we may make forward-looking statements regarding future events or the expected performance of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results could differ materially. For important factors that may cause actual results to differ from those contained in our forward-looking statements, please review our filings with the SEC.

The forward-looking statements made in this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, this presentation may not contain current or accurate information. We do not assume any obligation to update any forward-looking statements we may make. In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only and shall not be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionality described or to include any such feature or functionality in a future release.

Our Speakers

CELESTE TRETTO

Data Scientist

SARAH MOIR

Program Manager

Talk Contents

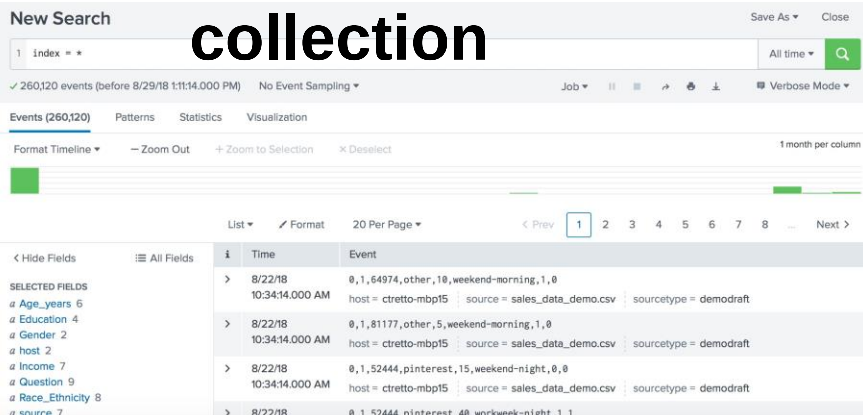
- What is in a machine learning model?
- How do machine learning models get biased?
- New and improved ways to spot bias
- How to address bias after you spot it

How algorithms get biased

What we covered last year

Components of a Machine Learning Model

Data collection



Feature engineering

purchase	clicks	on_sale	returning	time_period
0	10	1	1	weekend-morning
0	5	1	1	weekend-morning
0	15	0	1	weekend-night
0	40	1	1	workweek-night
0	5	0	0	workweek-evening
0	15	0	0	weekend-evening
0	15	0	0	weekend-afternoon
0	30	0	0	weekend-afternoon
0	25	0	1	weekend-evening
0	25	1	0	weekend-afternoon
0	30	1	1	workweek-afternoon

Algorithm

$$\hat{Y} = \omega X + \varepsilon$$

Model output

Example Machine Learning Model

Which universities are the best?

- Data Collection
  - N of professors / instructors
  - Research publications
  - Infrastructures
  - Classes
- Real Factors
  - Satisfaction
  - Personal growth
  - Career success
  - Happiness
- Model Proxy Features
  - Teacher / student ratio
  - SAT scores
  - Graduation rates
  - Employment rate
  - Reputation scores

Source: US News and World Report, “Weapons of Math Destruction” by Cathy O’Neil

It’s Easy to Introduce Bias

Data collection

Incomplete data
Biased data

Feature engineering

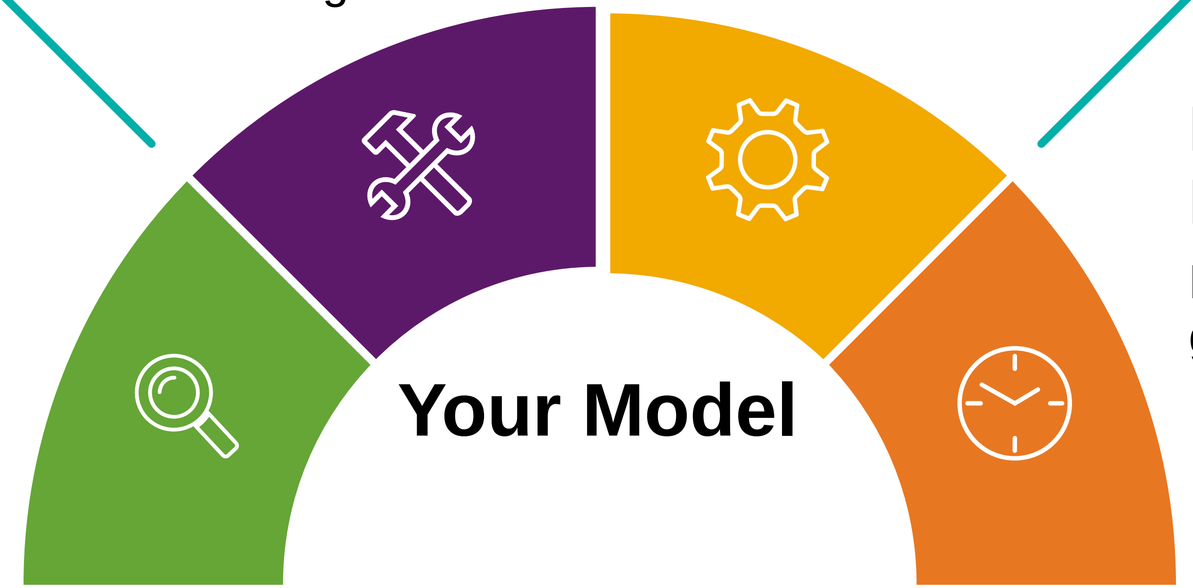
Subjective features
Incomplete features
Wrong features

Algorithm / Objective function

Opaque “black-boxes”
Standardized preferences

Decision-making / Feedback loops

Model outcomes have global impact



Key Takeaways

Recognizing bias in data requires everybody's best effort

1. Ask if the data is **representative**.
2. Ask if the data is **biased**.
3. Ask if the features are **accurate proxies**.
4. Ask if the goal of the model is **unbiased**.
5. Ask about the **implications** of the model results.

Spot bias in data

Methods to identify biased data

splunk> .conf18

Datasheets for Datasets

Keep Context with the Data

- Use and produce datasheets for datasets that you use and/or create
- Datasheets contain:
  - Why the dataset was created
  - What is in the dataset
  - How the data was collected
  - How the data was cleaned or pre-processed
  - Whether the dataset is maintained
- Helps you better identify biased data, or whether or not a specific dataset could lead to biased outcomes if used for a different purpose than the one for which it was originally used

Source: Gebru, Morgenstern, Vecchione, Wortman Vaughan, Wallach, Daume III, Crawford, 2018
https://arxiv.org/abs/1803.09010

Spot bias in models

Define Fair Model Outcomes

Define what fairness means

- **Fairness happens when all model components (data, features, algorithms) are not a function of a protected group**
- Model evaluation metrics should be similar among groups

- Remember the risk scores for recidivism we talked about last year?
  - Courts in the US use a mathematical “risk assessment” for individuals
  - Compare: model prediction (“High”, “Medium”, “Low” risk) vs real outcome (Conviction within 2 years)
  - How good was the model at predicting recidivism in general?
  - How good was the model at predicting recidivism by race?

Dashboard to Audit Algorithmic Bias

Demo time

- Evaluate the data for
  - Equal representation (representation balance of groups)
  - Equal real world outcomes (same distribution of real life outcomes)
- Evaluate the model for
  - Precision rate parity (true positives vs false positives)
  - Recall rate parity (true positives vs false negatives)
- Set a “fairness threshold”

- Code : https://github.com/ctretto/splunk-discriminatorybias

Biased Data

Demo time

Search | Datasets | Reports | Alerts | Dashboards | Search & Reporting

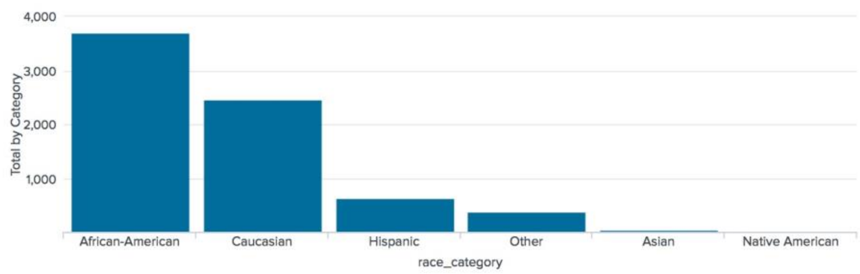
Classification Algorithm Evaluation

Edit | Export | ...

Data Source: index = * | eval pred = if(score_te
Audited Field: race_category
Reference Group: Caucasian
All time
Real Outcomes: is_recid
Predicted Outcomes: pred
Hide Filters

Representation by Category

Count of data points for each category



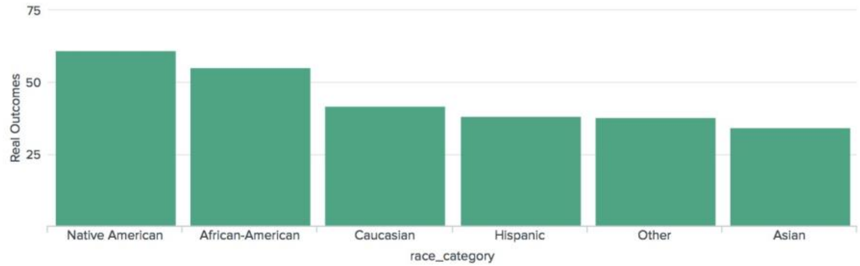
Representation Bias

Percentage difference with respect to reference category

race_category	Total by Category	Representation Difference
African-American	3696	50.61%
Caucasian	2454	0.00%
Hispanic	637	-74.04%
Other	377	-84.64%
Asian	32	-98.70%
Native American	18	-99.27%

Real Outcomes by Category

Percentage of real outcomes by category



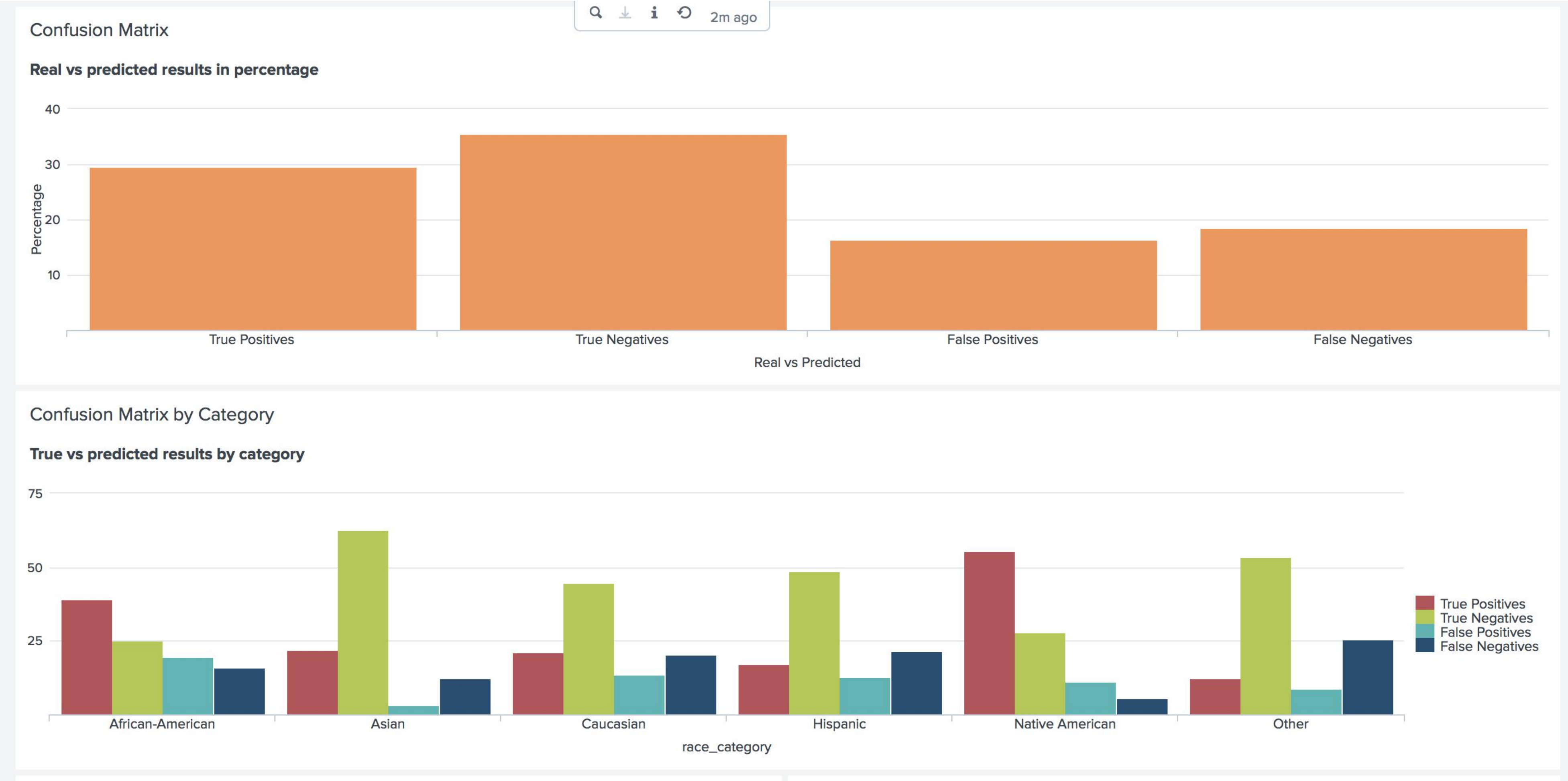
Real Outcomes Bias

Percentage difference with respect to reference category

race_category	Percentage by Category	Real Outcomes Difference
Native American	61.11	46.30%
African-American	55.09	31.89%
Caucasian	41.77	0.00%
Hispanic	38.46	-7.92%
Other	37.93	-9.19%
Asian	34.38	-17.69%

Poor Feature Engineering

Demo time

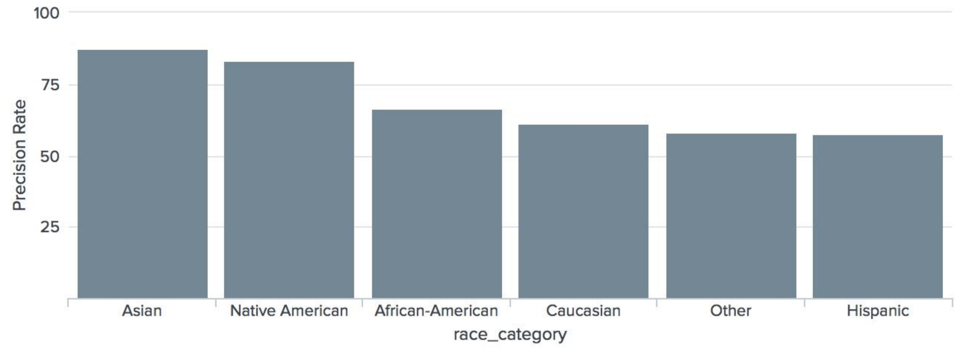


Leads to Poor Model Performance

Demo time

Precision Rate by Category

Out of all positive predictions, how often is the model correct



Precision Bias

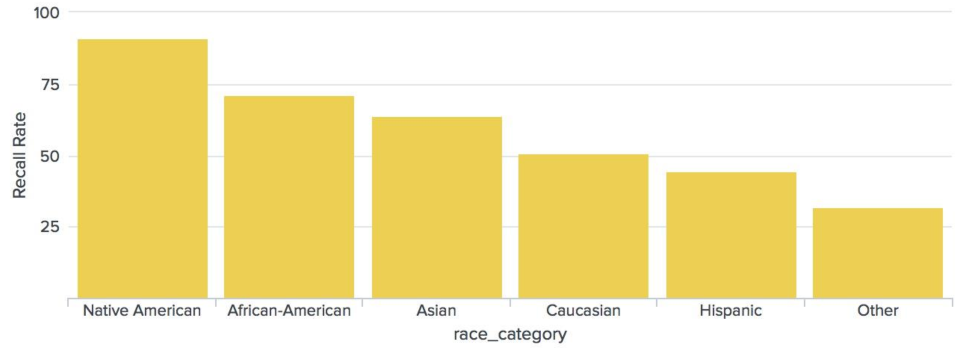
2m ago

Percentage difference in precision rate by category

race_category	Difference in Precision Rate
Asian	42.88%
Native American	36.07%
African-American	8.53%
Caucasian	0.00%
Other	-4.92%
Hispanic	-6.32%

Recall Rate by Category

Out of all true labels, how often is my model making a positive prediction



Recall Bias

Percentage difference in recall rate by category

race_category	Difference in Recall Rate
Native American	78.17%
African-American	39.10%
Asian	24.72%
Caucasian	0.00%
Hispanic	-12.81%
Other	-36.96%

Tools to Audit Algorithmic Bias

Online resources

- Aequitas: Open Source Bias Audit Toolkit from University of Chicago (https://dsapp.uchicago.edu/aequitas/)
  - Python tool similar to our dashboard
- TuringBox: Crowdsourcing model evaluation (https://turingbox.mit.edu/upload.html)
  - Still being developed

Fix your model

After you spot bias in your model, fix it

Fix a Biased Model

It's not simple but it is important

- Assumption: we want to avoid bias based on protected attributes like gender, race, age, etc.,
- Three approaches (that we'll talk about):
  - Consider fairness in your algorithm's objective function
  - Simulate multiple counterfactual worlds
  - Adversarial models

Different types of fairness

What do fair model outcomes look like?

- **Fair treatment**: model features are independent of protected attributes
  - Model cannot use protected attributes for prediction
  - Unrealistic assumption!
- **Fair impact**: model predictions are independent of protected attributes
  - Predictions for attribute = 0 are the same as predictions for attribute = 1
  - Equality of opportunity: Prob(correct_prediction and attribute = 0) = Prob(correct_prediction and attribute = 1)

Method 1: Consider Fairness in Objective Function

Include a fairness score in algorithm criteria

- Example: predict whether a purchase will be made
- Develop two models based on various features and train both to predict "Purchase"
  - Model1: some features are correlated with protected attributes (e.g., ZIP codes with race)
  - Model2: features are not correlated with protected attributes (e.g., returning customer)
- Based on the model outcomes, pick the "best" model
  - If "best" means "best at predicting purchase" we could pick a discriminatory model
  - Both Model1 and Model2 have "fair treatment" because the features are not directly reliant on protected attributes
  - But Model1 is still discriminating based on protected attributes

Method 1: Consider Fairness in Objective Function

Include a fairness score in algorithm criteria

- Proposal: include a “fairness” component in the objective function
- Think of it as two models in one
  - I want my features to be very good at predicting purchases, but
  - I do not want the quality of my prediction to be correlated with protected attributes

Model	“Traditional” model score	Unfairness penalization	Final score
Model 1	0.8	-0.25	0.55
Model 2	0.7	-0.1	**0.6**

- Drawback:
  - Prediction power of Model2 is not as good
  - The quality of the model depends on how much discrimination bias is present in the data

Source: M.B. Zafar, Valera, Gomes Rodriguez, Gummadi 2015

https://arxiv.org/abs/1507.05259

Method 1: Takeaways

Include a fairness score in algorithm criteria

- Think of avoiding bias as a feature selection / model comparison process
- Compare models, play with your features
- Predict sensitive attributes using your model's features
- Use the results from the Bias Audit Dashboard to build an unfairness penalization score

Method 2: Assess Results in a Counterfactual World

Addresses historical bias present in data

- A decision is fair towards an individual if it's the same in the actual world and in a counterfactual world
- Example: determine law school success given SAT scores and GPA
- Proposal: add unknown social biases to the model
- Multi-step process:
  - Step 1: Simulate numerous versions of a socially biased world.
  - Step 2: Based on those simulations, create a "knowledge" factor that cannot be observed but normalizes the social biases across those simulated worlds.
  - Step 3: Predict law school grades using SAT scores, GPA, protected attributes, and the knowledge factor

Source: Kusner, Loftus, Russel, Silva 2018

https://arxiv.org/abs/1703.06856

Method 2: Takeaways

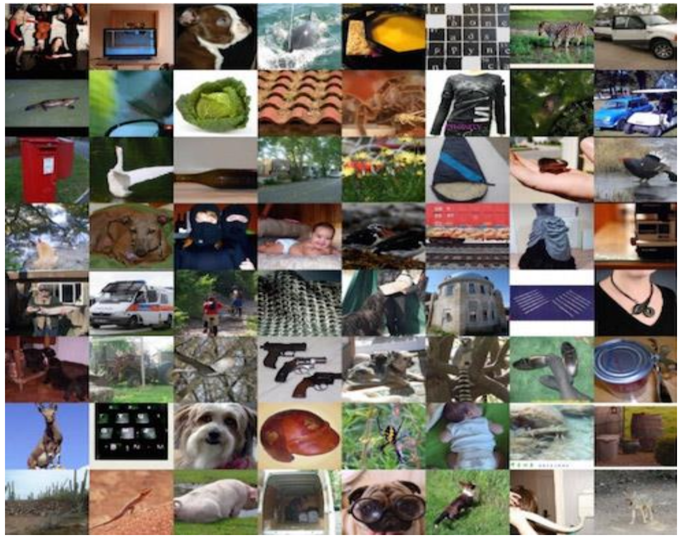
Addresses historical bias present in data

- Can you build better features?
- Can you research what are the social biases that are present in the world you are modeling?
- Present your model findings together with other sources of information

Method 3: Generative Adversarial Models

Generative vs. discriminative models

- Generative algorithm models generate data with the same structure as original data
- Generative Adversarial Networks are a class of neural networks
- Generative Adversarial Networks (GANs) are composed of
  - Generator: generates data
  - Discriminator: tells the difference between real data and generated data

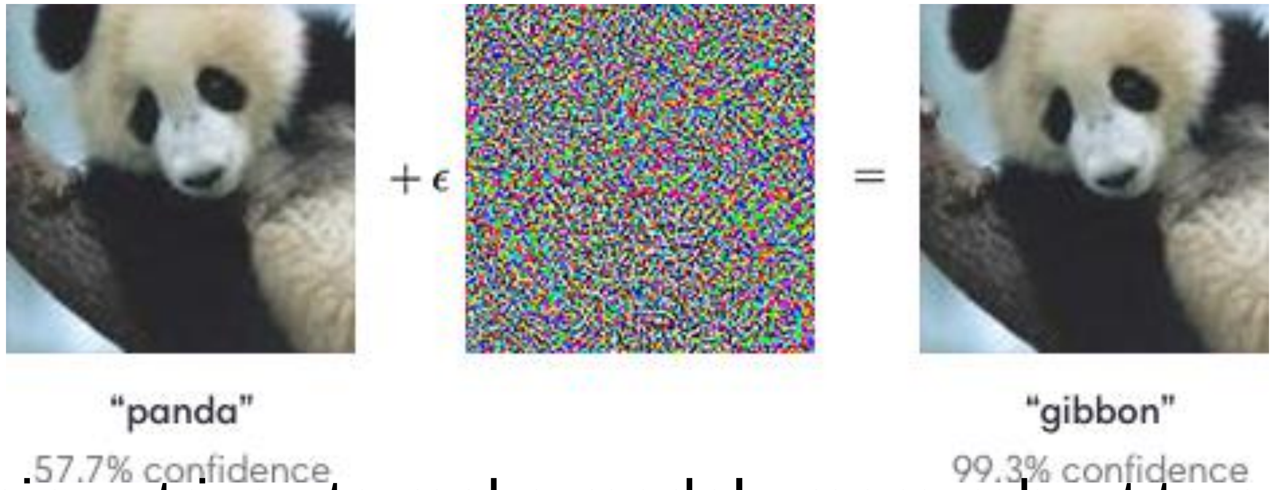


Learn mode: openai.com

Method 3: Generative Adversarial Models

Generative vs. discriminative models

- Adversarial models are also a way to corrupt the inputs of a model to intentionally pollute the results



- Adversarial learning strives to make models more robust to noise in the data
- What if bias was the noisy component?

Source: Zhang, Lemoine, Mitchell 2018 https://arxiv.org/abs/1801.07593

Source: Xu, Zhang, Yuan, Wu 2018 https://arxiv.org/pdf/1805.11202.pdf

Method 3: Takeaways

Include a fairness score in algorithm criteria

- Can you simulate unbiased data?
- Can you resample your data in order to avoid some of the biases?
- Compare model results with different datasets, both real and simulated
- GANs are available with Keras and TensorFlow

Conclusion

Key Takeaways

Pandas are not gibbons

1. Machine learning is **not doomed**.
2. Determine what **types of bias** you want to address.
3. Write datasheets for datasets to **prevent potential data bias**.
4. Use automated tools to **identify model bias**.
5. Use the available **methods to reduce bias**.

Thank You!

Don't forget to rate this session in the .conf18 mobile app

.conf18

splunk>