Machine Learning Bias

AbdulMajedRaja RS

Outline

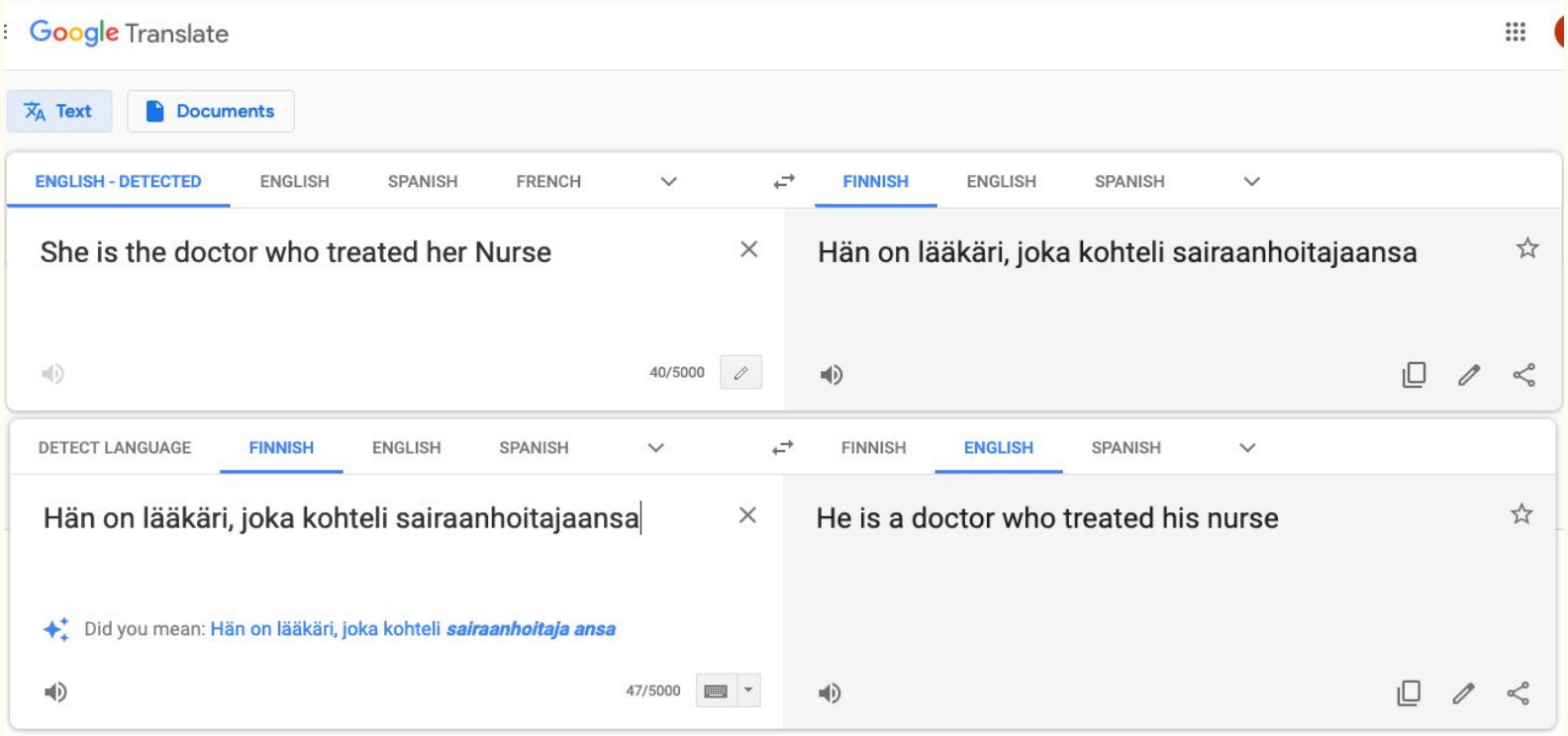
- Recognizing the Problem
- What's Machine Learning Bias?
- Definition of "Fairness"
- Interpretable Machine Learning
- Case Studies

Thoughts?

What if I told you Computers can lie?

Would you believe me?

Biased-Google Translation at Work

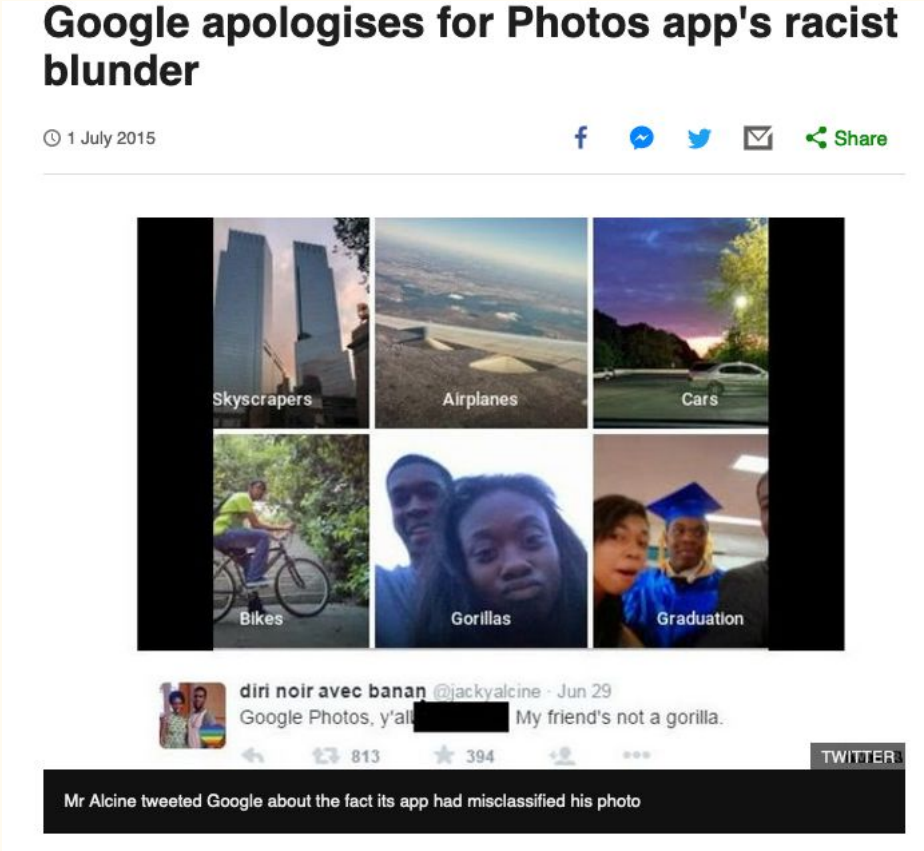


The Problem - Samples

But Wait, Why is this concerning?

After all, This is just Google Translate

Biased-Google Photos App at Work



Perhaps, That’s just Google.

Two instances can account for the entire industry, Huh?

Microsoft's super-cool Teen Tweeting Bot Tay

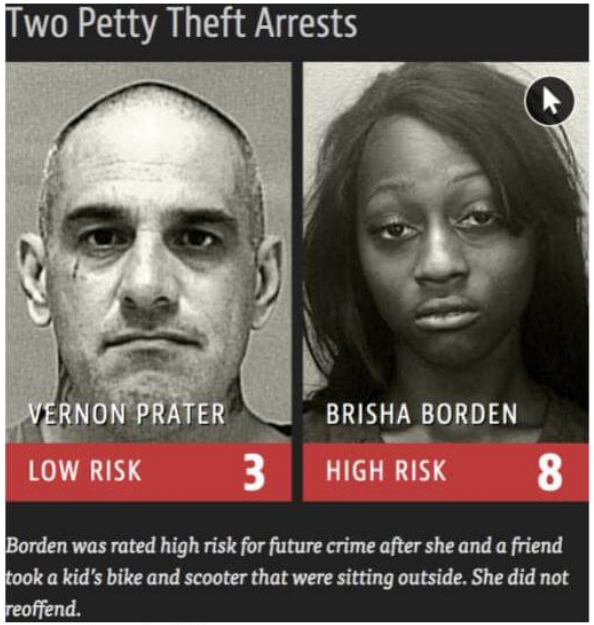
Much more!

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016



Oops, Got it!

There, definitely, is Bias!
What's next?

ML Bias - What

What's Machine Learning Bias?

A Machine Learning Algorithm being
"**unfair**" with its Predictions

A Machine Learning Algorithm missing
"**Fairness**"

ML Bias - (un)Fairness

Disclaimer

No Common Consensus / Standard definition of Fairness

ML Bias - un(Fairness)

- Group Fairness
- Individual Fairness

ML Bias - Causes

ML Bias - Causes

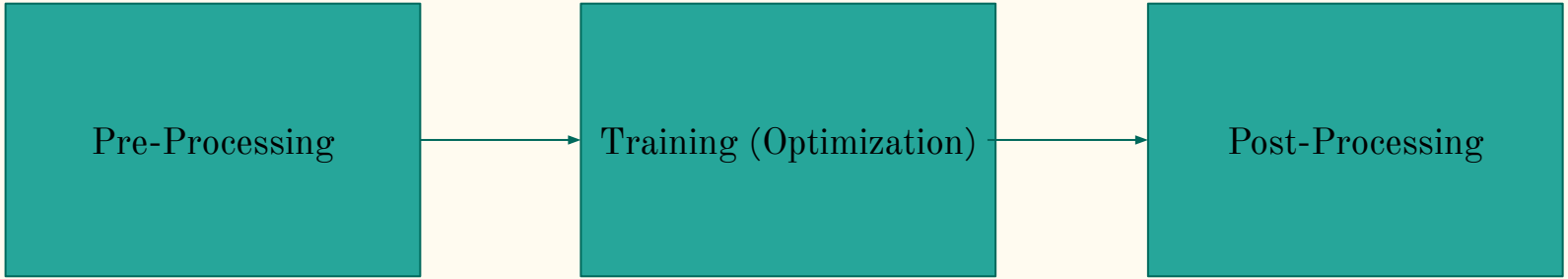
- Skewed sample
- Tainted examples
- Limited features
- Sample size disparity
- Proxies

ML Bias - Mitigate

Mitigation

Also means, Improving Fairness

ML Bias - Improving Fairness



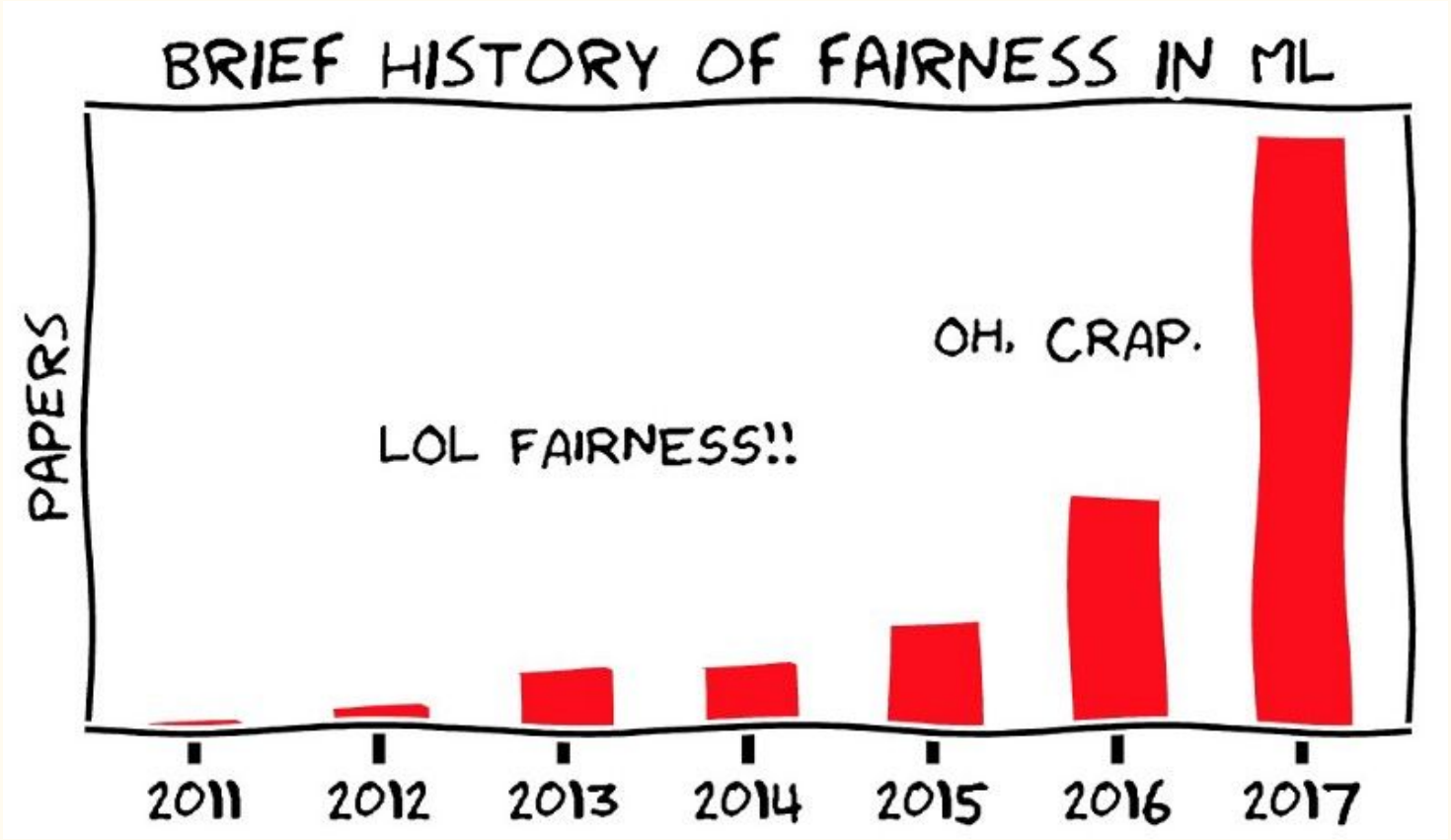
Learn a New Representation - Free from Sensitive Variable Yet, preserving the Information

Add a constraint or a regularization term

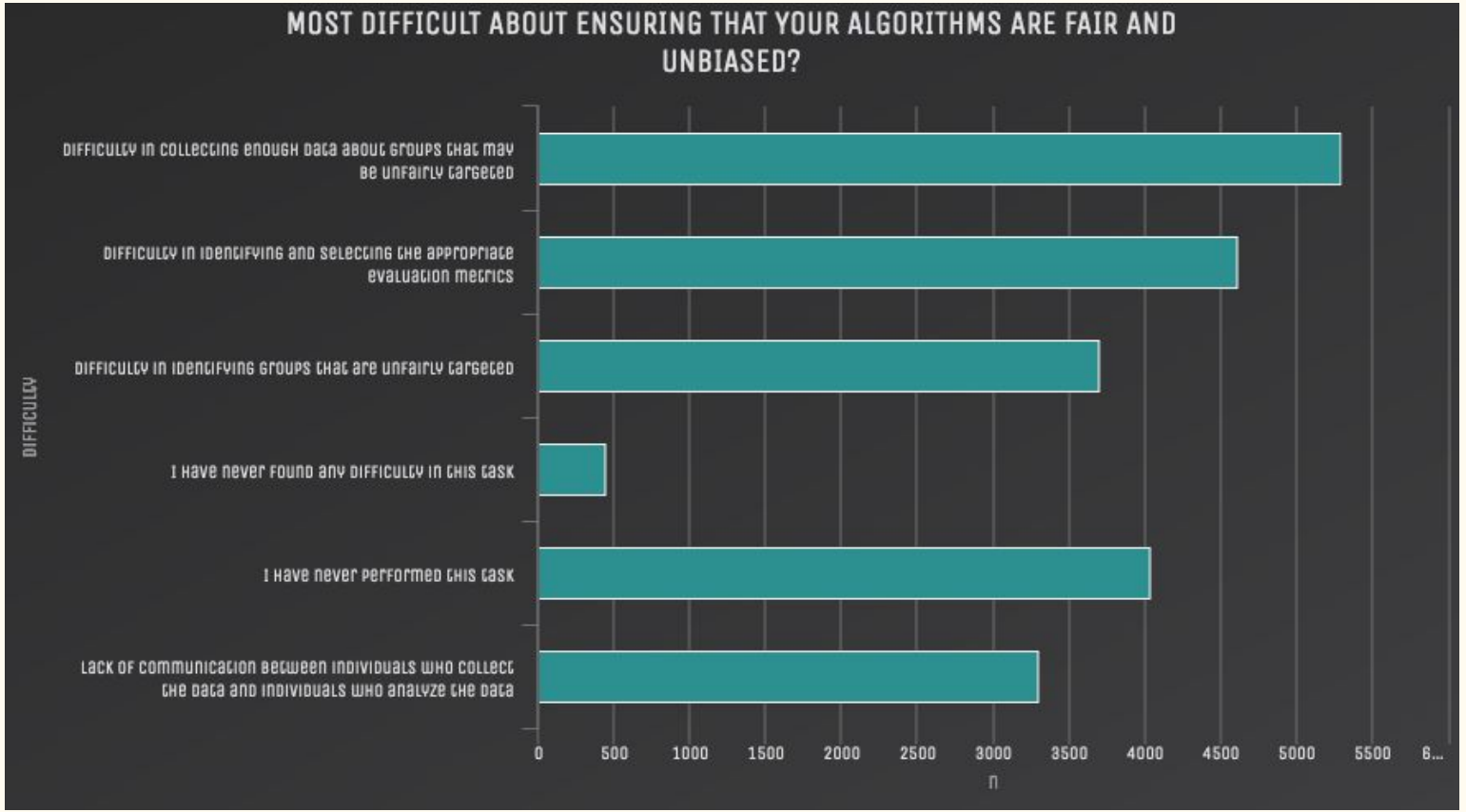
Find a proper threshold using the original score function

ML Bias - Happening

Mention of ML Fairness in Research Papers

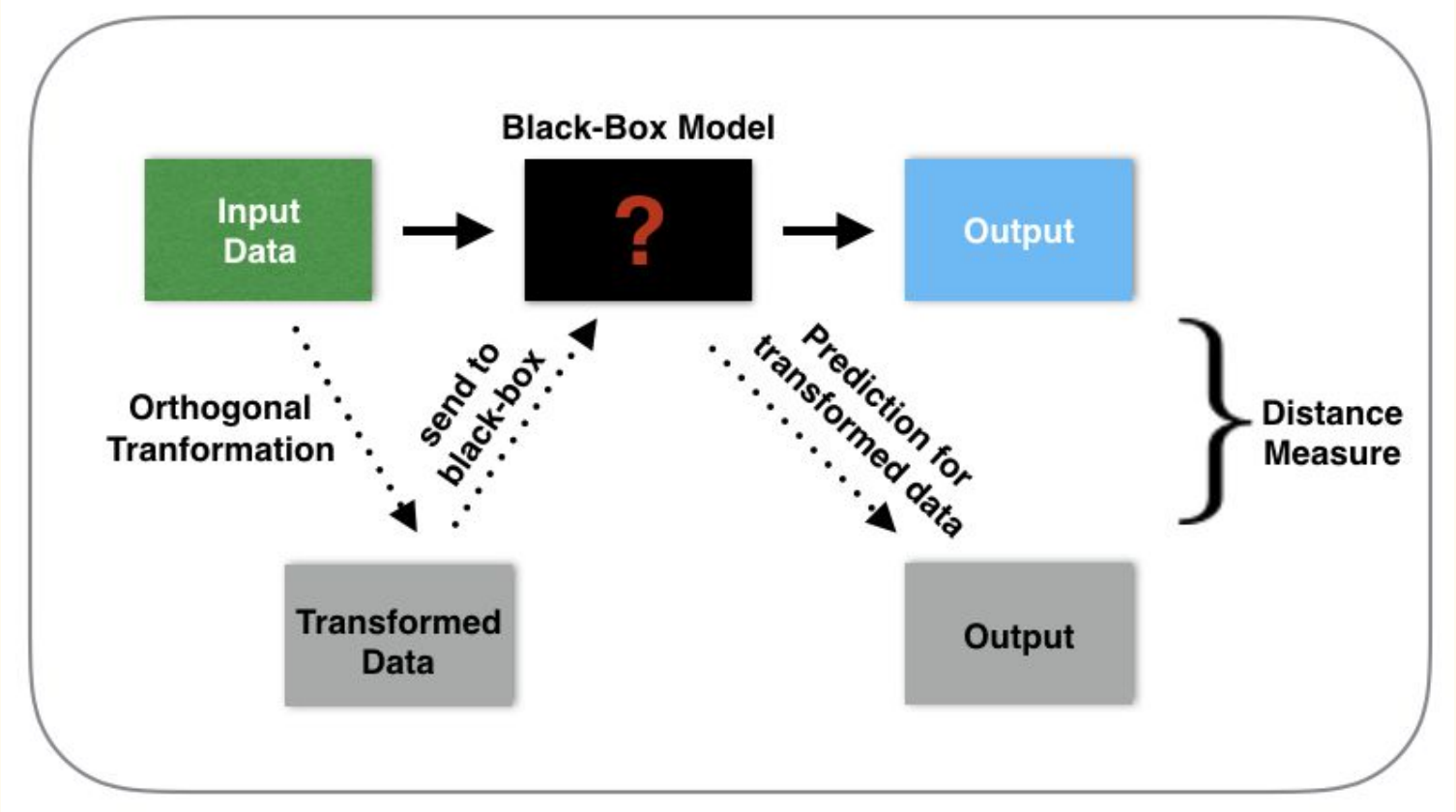


Difficulties in ensuring ML Algorithm is unbiased



Interpretable Machine Learning

Today - Modelling Architecture



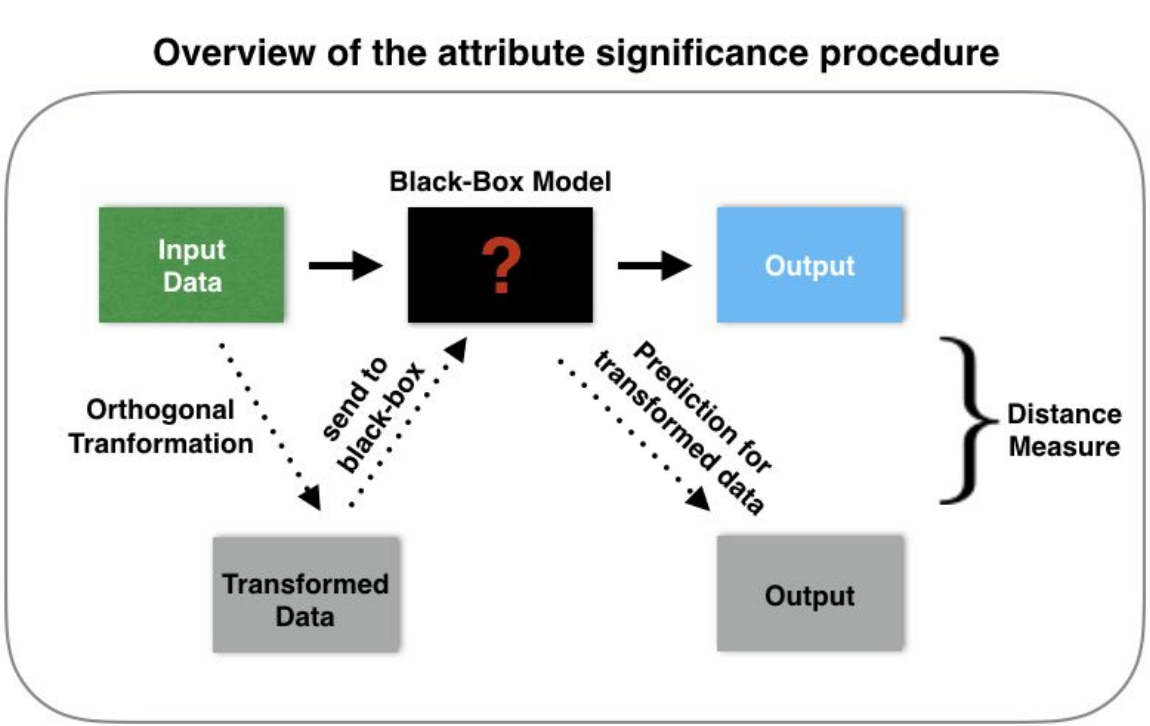
IML - Definition

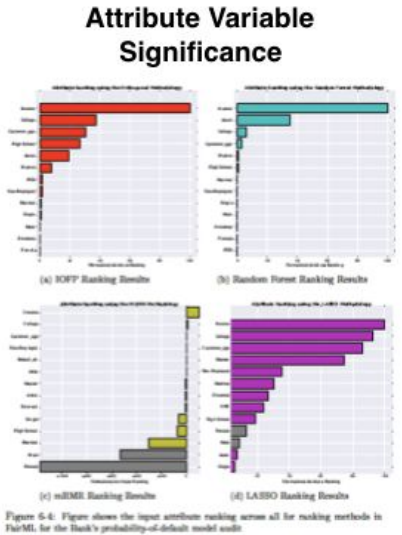
Interpretable Machine Learning refers to methods and models that make the behavior and **predictions of machine learning systems understandable to humans**.

IML - Benefits

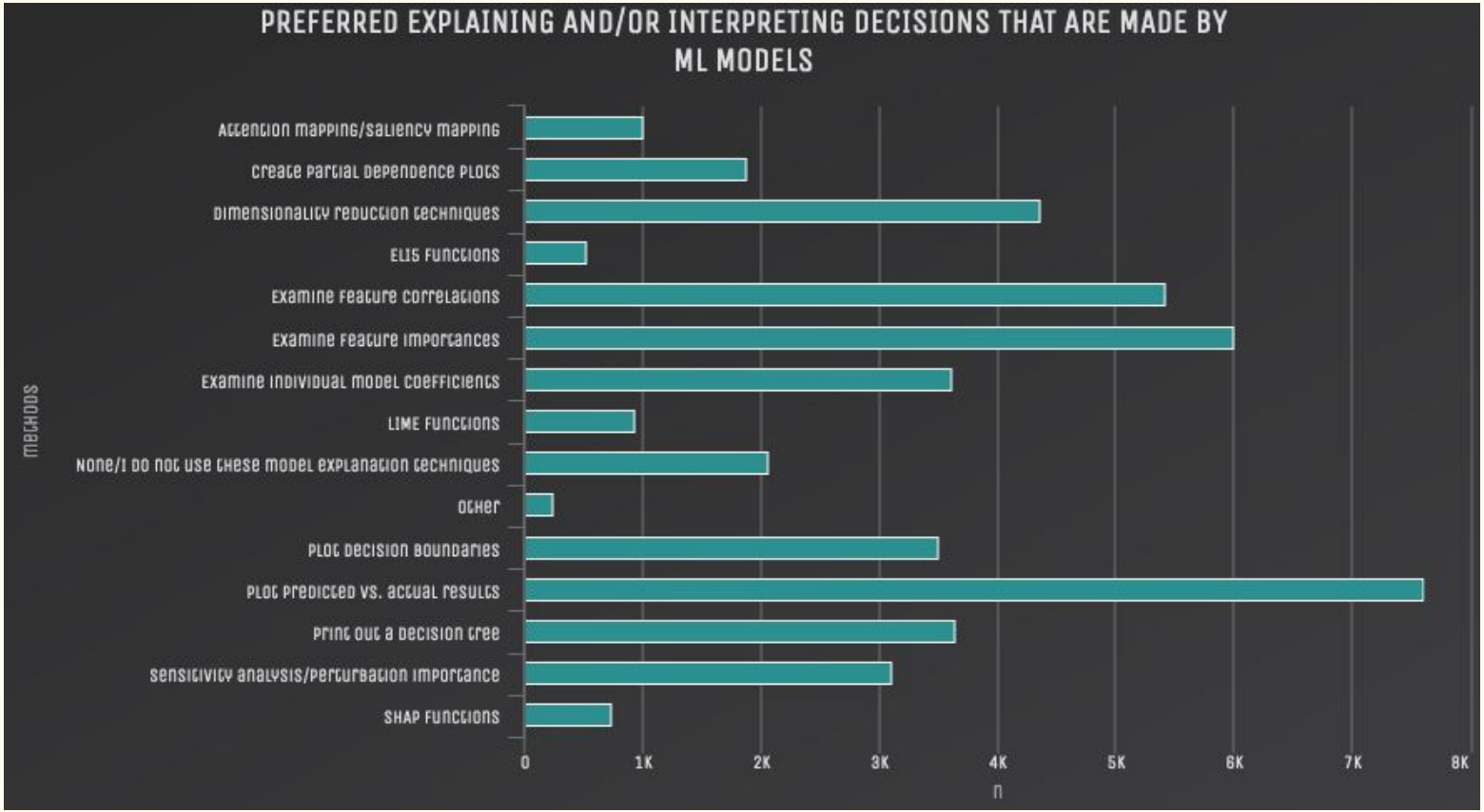
- **Fairness**: Ensuring that predictions are unbiased and do not implicitly or explicitly discriminate against protected groups. An interpretable model can tell you why it has decided that a certain person should not get a loan, and it becomes easier for a human to judge whether the decision is based on a learned demographic (e.g. racial) bias.
- **Privacy**: Ensuring that sensitive information in the data is protected.
- **Reliability or Robustness**: Ensuring that small changes in the input do not lead to large changes in the prediction.
- **Causality**: Check that only causal relationships are picked up.
- **Trust**: It is easier for humans to trust a system that explains its decisions compared to a black box.

Modelling Architecture - with IML





Preferred Explaining - Model Interpretation



Case Studies

Attrition Prediction - People Analytics

Objective:

- Building a Machine Learning Solution to Predict Employee Attrition in the Organization for the next Quarter

Organization: Confidential

Data used:

- Demographic data
- Compensation data
- Promotion data
- Reward & recognition Data

Attrition Prediction - People Analytics

Final Model:

- Ensemble of Bagging (RandomForest) and Boosting (xgboost) with weighted Average

Accuracy: Acceptable

All Good?

Can we go ahead and Productionize?

But, Wait!

Attrition Prediction - People Analytics

Interpreting the Model:

- Variable Importance Plot - for unboxing the Blackbox methods

Result:

- **Maternity Leave** (x) is one of the most **important Variable to Attrition** (y)

Machine Learning Ethics?

Attrition Prediction - People Analytics

Implications with Proceeding with this Model

- Female Employees taking Maternity Leave would be suspected of Leaving the Job soon
- Future Hiring of Married Female Employees would be scrutinized

Attrition Prediction - People Analytics

Implications with Proceeding with this Model

- Female Employees taking Maternity Leave would be suspected of Leaving the Job soon
- Future Hiring of Married Female Employees would be scrutinized thus having an impact in such hires

Implications of NOT Proceeding with this Model

- Accuracy would be impacted as the current score has already hit acceptable leve
- Less trust on Data Science teams by other Cross-Functional Teams

Attrition Prediction - People Analytics

Result

- Retrained the Model with `Maternity Leave` made a `Protected Attributed` and made `**unaware**` to the Model during the Training
- Thus, Newly built model excludes the Sensitive Variable (`Maternity Leave`) that lead to Bias against a particular segment (`Female & Married`)

Impact

- Reduction in Model Accuracy Score
- But, Job Delivered to the HR Department with a Model of No Obvious Bias in it

Attrition Prediction - People Analytics

Lessons Learnt

- Unlike the obvious presence of Bias from Data being transferred to the Model, In this case, there's no Bias (as such) in the Data
- But the Model during the Training (Feature Engineering) learnt which leads to Bias
- Mostly, It comes down to Trade-off between Accuracy and Responsible Data Science
- Better techniques, just other than `unaware` could have been used to minimize the accuracy loss
- Machine Learning Ethics Matter to be built something that's fair to everyone

References

- https://developers.google.com/machine-learning/fairness-overview/
- https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb
- https://www.youtube.com/watch?v=fMym_BKWQzk
- https://www.kaggle.com/nulldata/ml-bias-iml-perspective-recommendation#media-coverage-about-bias-iml
- Doshi-Velez, Finale, and Been Kim. “Towards a rigorous science of interpretable machine learning,” no. Ml: 1–13. http://arxiv.org/abs/1702.08608 (2017)
- https://christophm.github.io/interpretable-ml-book/
- https://github.com/adebayoj/fairml/
- https://arxiv.org/pdf/1904.05233v1.pdf

Thank you!