Machine Learning Bias and Fairness

Dr. Brian Wright

Practice of Data Science

Machine Learning Bias and Fairness

- What does **Machine Learning Bias** mean?
- The term "bias" as it applies to machine learning was introduced by Tom Mitchell in 1980.
  - "any basis for choosing one generalization (hypothesis) over another, other than strict consistency with the observed training instances"

- What this really means is essentially that every learning algorithm must accept a certain level of bias in order to generalize beyond the data it is provided.

- Bias is more or less a requirement

Dietterich, T., & Kong, E.B. (1995). Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree Algorithms, University of Oregon.

Machine Learning Bias and Fairness

- The question then is when does bias manifest as socially unaccepted?

- This form of "bias" is a result of **human generated bias** that is inherit in our society and replicated into datasets and learned into decision algorithms.

- **Cognitive bias** – Foundation elements of the field of behavioral economics, suggests that humans do not make rational decision but are instead influenced by their environments.
  - This results in perceptual distortion, inaccurate judgement, illogical interpretation or simply acting irrational. All of which can manifest in data. [2]

Kahneman, D.; Tversky, A. (1972). "Subjective probability: A judgment of representativeness". Cognitive Psychology.

Machine Learning Bias and Fairness

- Two quick examples of ML Bias at work

- Methods for identifying potential Bias
  - Discussion on ML Evaluation Metrics

Machine Learning Bias and Fairness

Examples

Machine Learning Bias and Fairness

- Examples:
  - Google's Cloud Natural Language API was launched in 2016. In fall of 2017 Andrew Thompson from Motherboard Inc. experimented with the tool and discovered bias results.
  - "I'm Christian" had positive results but, "I'm a Jew" and "I'm a gay black woman" resulted in negative results.
  - Trained on news and social media data

Kahneman, D.; Tversky, A. (1972). "Subjective probability: A judgment of representativeness". Cognitive Psychology.

Machine Learning Bias and Fairness

From Google: “We dedicate a lot of effort to making sure the NLP API avoids bias, but we don’t always get it right. This is an example of one of those times, and we are **sorry**.” ☹

Machine Learning Bias and Fairness

- Example:
  - Amazon Recruiting Engine
    - Used an AI tool to give candidates a score from one to five stars based on resumes.
    - "They (Amazon) literally wanted it to be an engine where I'm going to give you 100 resumes, and it will spit out the top five, and we'll hire those."
    - The problem was that data was used from 10 years of hiring that was mostly men, so the ML tool systematically scored women lower
    - Amazon just this year abandoned the project because it couldn't be proven reliable.

Machine Learning Bias and Fairness

- So if Amazon and Google can't get this right where does that leave us moving forward?

- To a certain extent we can never eliminate the **human/cognitive bias** present in our society, as a consequence training data may always be corrupt.

- Until a time in which machine learning improves to the point that these biases can be accounted for we need to make thoughtful decisions about how we use these tools, lucky some techniques are emerging.

Assessing Fairness of ML Models

Machine Learning Bias and Fairness

- Defining Fairness – The field is starting to mature around some central approaches to assessing bias and fairness of machine learning algorithms
  - Fairness (non-bias) - outcomes of a machine learning can be mathematically proven to treat the protected classes equally – **group bias**
    - Meaning the errors of a model are similar across these groups. These groups usually represent "protected classes" such as gender or race.
  - Keep in mind that there may be **statistical variations** in outcomes between these groups already, what we want is the machine to take these into account but not unfairly penalize the protected class variables in comparison to those that have similar profiles that don't represent a protected class – **individual fairness**

Machine Learning Bias and Social Science Methods

- In assessing whether our models are "fair" we can use several rather simple assessment approaches
  - Demographic Parity
  - Equality of Odds
  - Equality of Opportunity

- But…before we dig into these, we need to discuss evaluations of ML models, specifically confusion matrices and AUC/ROC curves.

Machine Learning Evaluation

Machine Learning Bias and Social Science Methods

- Confusion Matrix, ROC (receiver operating curve) and AUC (area under the curve) are very common approaches for measuring the performance of classification models
- Classification models output percentages that an individual input will belong to a specific class, usually a 1 or 0, with one being a positive attribute.
  - Likelihood of email spam/fraud is an example. The higher the model percentage prediction on any one email the higher chance it is fraud.
- Essentially both measure the misclassification error rate associated with your model
- A Confusion Matrix is a good tool for understanding how accurate you model is classifying and is used to build ROC

Machine Learning Bias and Social Science Methods

- Let's use intruder/fraud detection as an example
- Say we have 135 emails entering our system and we are trying to detect whether they are fraudulent or not
  - We use lots of criteria – source, subject, if they came from a prince…
- Generate probability measures as a result for a tree-based classifier to determine the likelihood that any one of these emails is fraudulent
- The cutoff point that is predetermined in the tree (and is a universal standard) is 50% but can be modified as an input if needed

Machine Learning Bias and Social Science Methods

- Below is are the results of our model in a **Confusion Matrix**. They center on the positive and negative classifications in sub-categories of true and false positive.
- Keep in mind we know because of the labels, what is fraud and not, so we can measure how good the model is classifying.
- Both true negative and true positive are good, false negative and false positive are errors.

1 = Fraud/Spam 0 = Not Fraud/Spam		Predicted Class	
		Positive Fraud Pred (1)	Negative Not Fraud Pred (0)
Actual Class	Positive Fraud Actual (1)	True Positive 10	False Negatives 22
	Negative Not Fraud Actual (0)	False Positives 7	True Negative 96

Machine Learning Bias and Social Science Methods

- Let's consider the extremes: what if we set the threshold to 0?
  - Means that everything is captured as Fraud and no ever gets an email again!

1 = Fraud/Spam 0 = Not Fraud/Spam		Predicted Class	
		Positive Fraud Pred (1)	Negative Not Fraud Pred (0)
Actual Class	Positive Fraud Actual (1)	True Positive 32	False Negatives 0
	Negative Not Fraud Actual (0)	False Positives 103	True Negative 0

Machine Learning Bias and Social Science Methods

- Let's consider the other extreme: what if we set the threshold to 100?
  - Means nothing is fraud and now everyone is getting rich off of Arabian princes

1 = Fraud/Spam 0 = Not Fraud/Spam		Predicted Class	
		Positive Fraud Pred (1)	Negative Not Fraud Pred (0)
Actual Class	Positive Fraud Actual (1)	True Positive 0	False Negatives 32
	Negative Not Fraud Actual (0)	False Positives 0	True Negative 103

Machine Learning Bias and Social Science Methods

- We can further assess our model by generating classification rates:
  - True Positive Rate (TPR) (Sensitivity) = TP/(TP+FN) = 10/(10+22) = .31
    - % of fraud correctly labeled as fraud
  - False Positive Rate (FPR) (1- Specificity) = FP/(FP+TN) = 7/(7+96) = .06
    - % of emails labelled not fraud that were false positives

1 = Fraud/Spam 0 = Not Fraud/Spam		Predicted Class	
		Positive Fraud Pred (1)	Negative Not Fraud Pred (0)
Actual Class	Positive Fraud Actual (1)	True Positive 10	False Negatives 22
	Negative Not Fraud Actual (0)	False Positives 7	True Negative 96

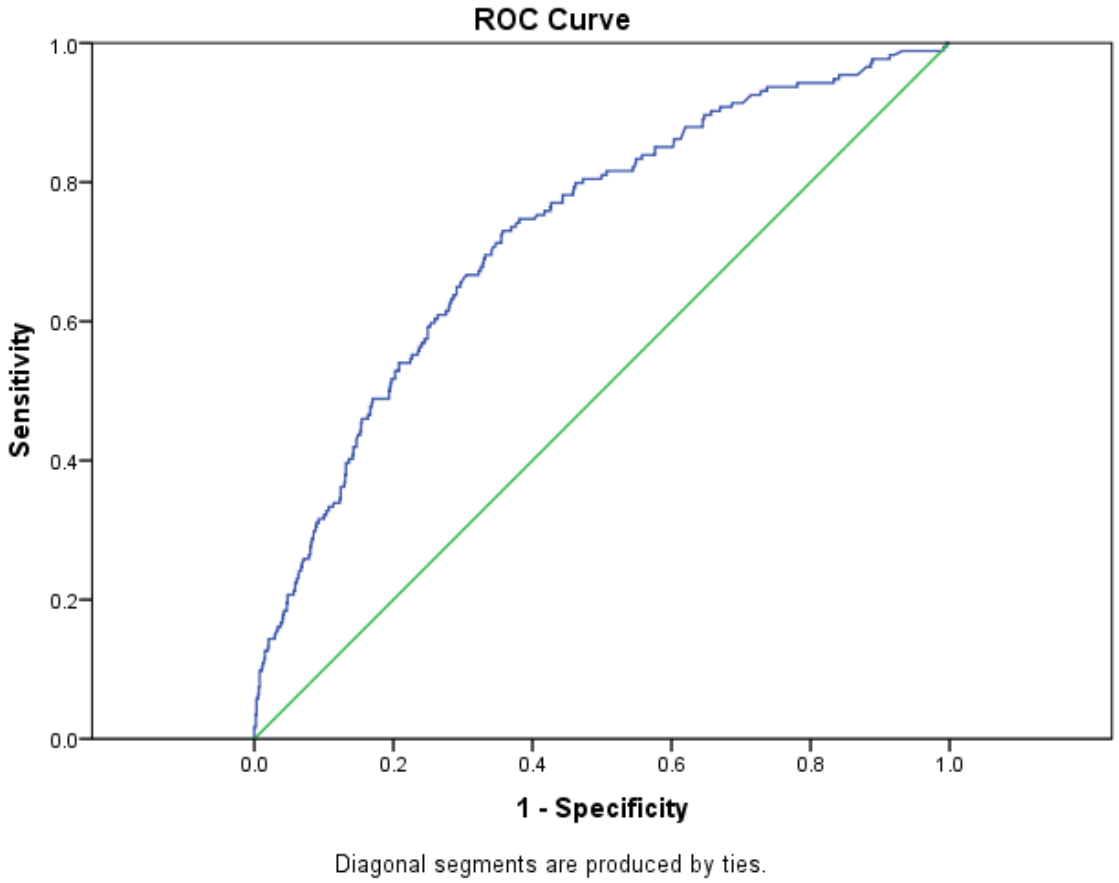
Machine Learning Bias and Social Science Methods

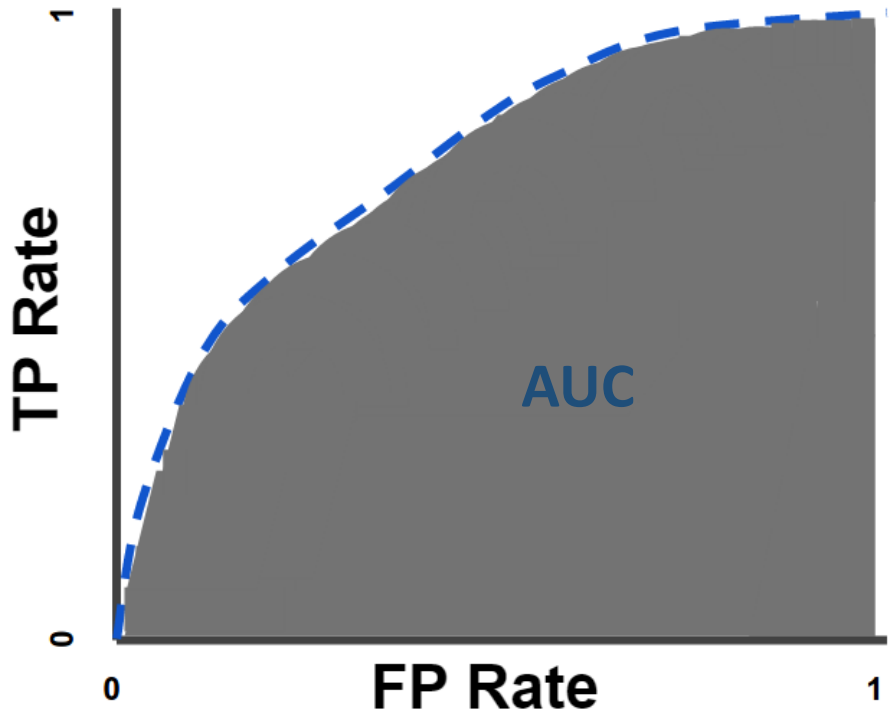
- These two data points can used to begin to develop a Receiver Operating Characteristic Curve or ROC curve
  - True Positive Rate (TPR) = 10/(10+22) = .31 = y-axis
  - False Positive Rate (FPR) = 7/(7+96) = .06 = x-axis

1 = Fraud/Spam 0 = Not Fraud/Spam		Predicted Class	
		Positive Fraud Pred (1)	Negative Not Fraud Pred (0)
Actual Class	Positive Fraud Actual (1)	True Positive 10	False Negatives 22
	Negative Not Fraud Actual (0)	False Positives 7	True Negative 96

Machine Learning Bias and Social Science Methods

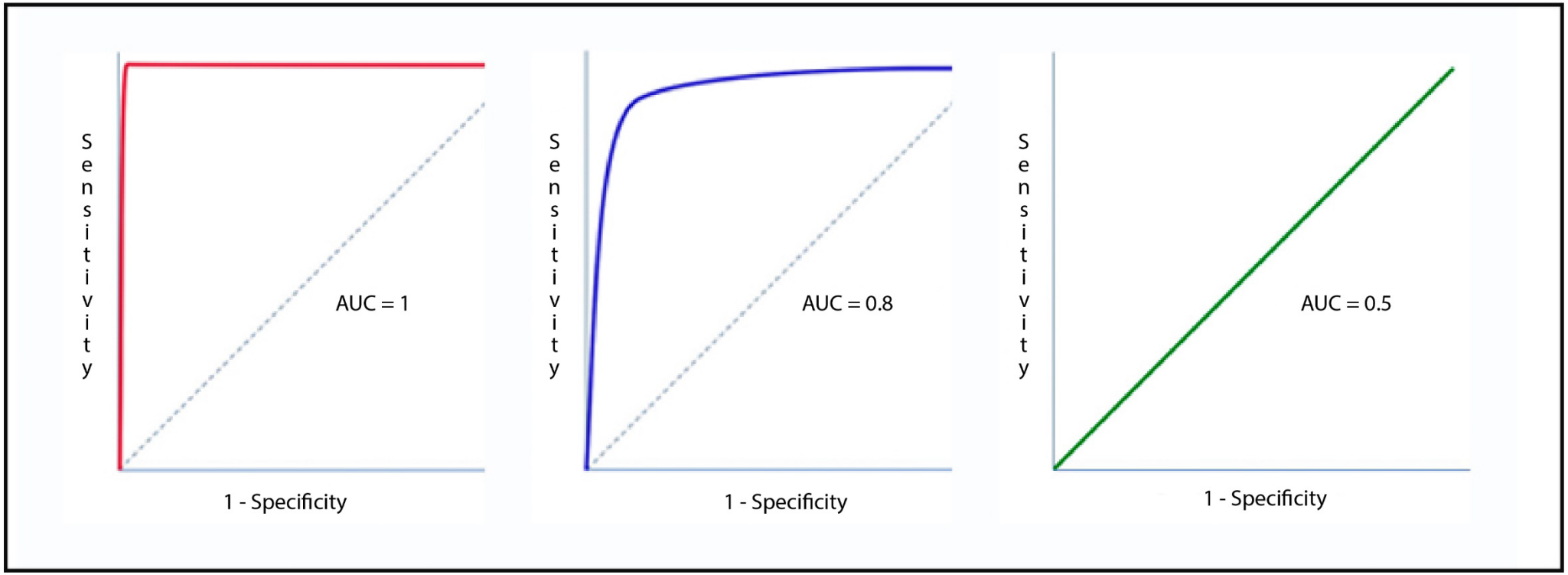
- ROC curve is essentially a graphical representation of the adjusted threshold values of the confusion matrix, below are two examples





Machine Learning Bias and Social Science Methods

- ROC curve generates the area under the curve as a percentage of the total graph under the curve.



Machine Learning Bias and Fairness

- These metrics can be used to assess the fairness of the machine learning models

- We will review these topics again later in the semester, as there's much more to be learning but having a basic understand will help when we walk through the fairness formulas.

Fairness Evaluation Methods

Machine Learning Bias and Fairness

- Demographic (proportional) Parity – Works to determine if each of the sub-groups (protected classes) have the same proportion in the positive category.
  - As an example, assume we develop a model of assessing worthiness for getting a loan. If the same relative percentages of all race in the model where deemed worthy of getting a loan, then is measure would be satisfied.
  - However, this is not overly useful. It's basically over simplified and could limit our ability to understand more discrete issues with the performance of the machine. For instance false positive and true positive rates could vary between classes with out our knowledge.

Machine Learning Bias and Fairness

- Equality of Odds– Works to determine if predictions are conditional independent from the protected class. Sensitivity – TP/TP + FN
  - It is a measure of if the true positive rate and false positive rate are the same for the various protected classes across different threshold levels.
  - This is basically ROCs for each of the sub-groups. How well does the model preform if the class are broken apart and analyzed.
  - Below A is the protected class with a binary classifier, for this to be true the odds of being placed in either 0 or 1 should be independent of A.

$$\Pr\left\{\widehat{Y}=1 \mid A=0, Y=y\right\}=\Pr\left\{\widehat{Y}=1 \mid A=1, Y=y\right\}, \quad y \in \{0,1\}^{1}$$

[1] Hardit, M, Price, E, & Srebro, N. (2016). *Equality of Opportunity in Supervised Learning.* NIPS, Barcelona, Spain.

Machine Learning Bias and Fairness

- Equal Opportunity– Extension of Equality of Odds but only focuses on the positive attribute, typically the classifier of 1 in the binary case. (R – Predicted Rate Parity)

  Precision – TP/TP+FP

  - We want the true positive rate to be the same for each protected class

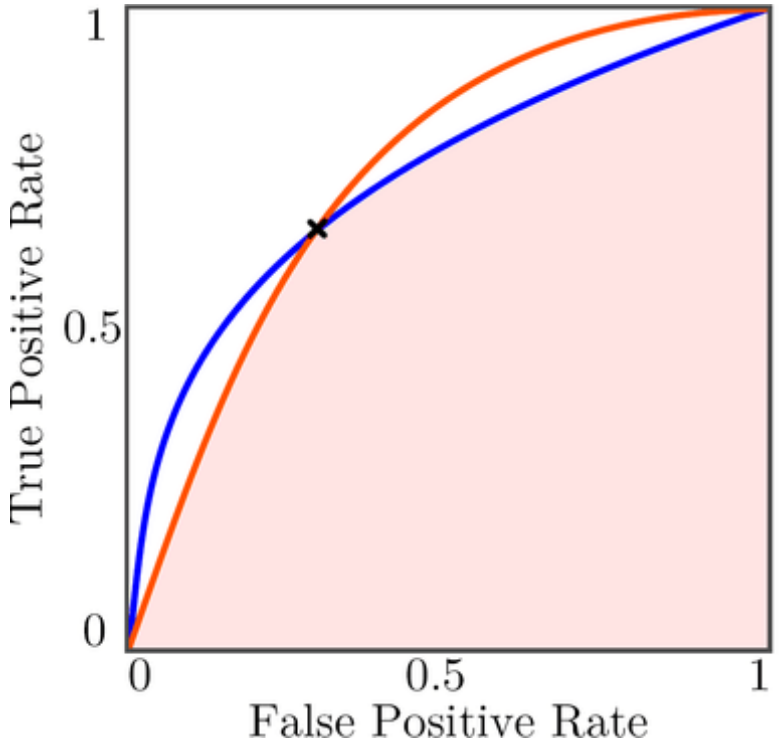
  - Below *A* is the protected class with a binary classifier, for this to be true the odds of being classified as 1 should be independent of *A*.

$$\{\hat{Y} = 1 \mid A = 0, Y = 1\} = \Pr\{\hat{Y} = 1 \mid A = 1, Y = 1\}$$

[1] Hardit, M, Price, E, & Srebro, N. (2016). *Equality of Opportunity in Supervised Learning.* NIPS, Barcelona, Spain.

Machine Learning Bias and Fairness

- The idea behind these methods is to find the ideal threshold that allows the equations to be satisfied.



[1] Hardit, M, Price, E, & Srebro, N. (2016). *Equality of Opportunity in Supervised Learning.* NIPS, Barcelona, Spain.

Machine Learning Bias and Fairness

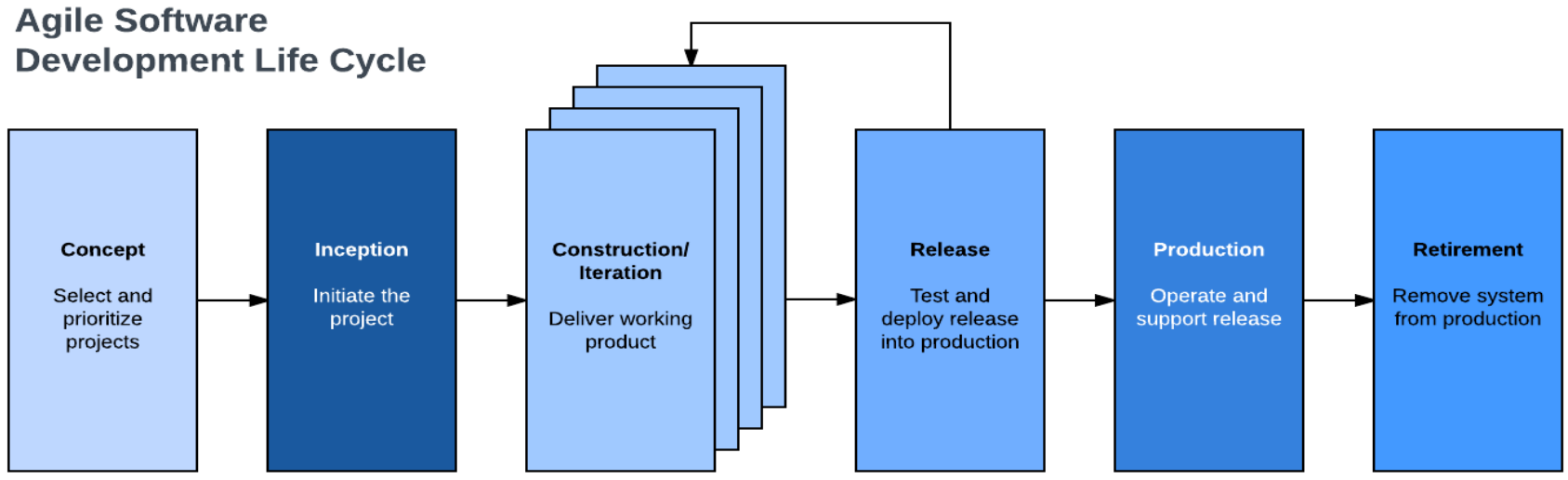
- Threshold shift to satisfy any of these conditions can be difficult and sometimes not feasible given the data at hand.

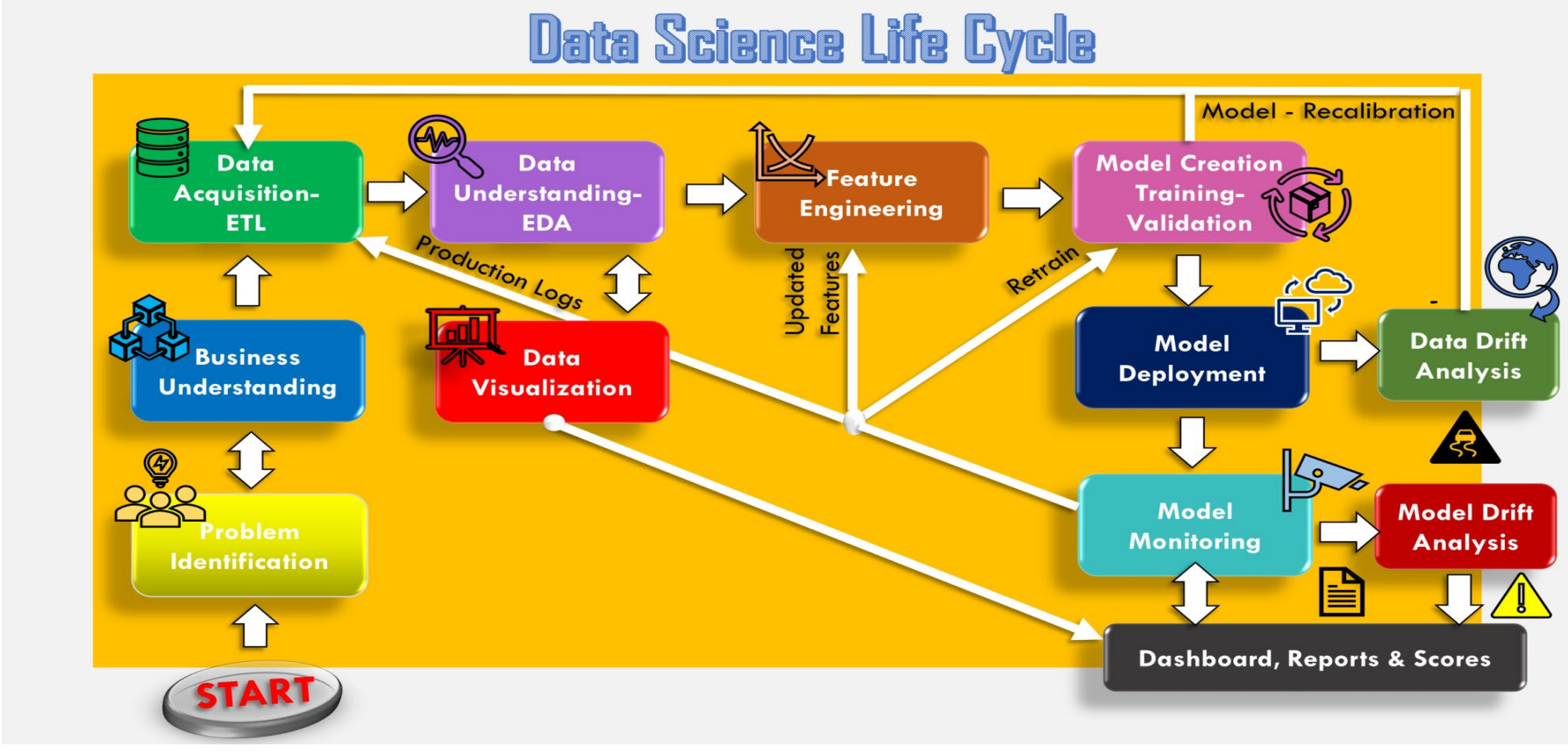
- Often, we can control for these differences prior to developing our algorithms through more complicated methods that work to balance the differences between classes.
  - You can think of these as methods weighting the variables ahead of time to avoid any potential bias.

Machine Learning Bias and Fairness

- Make sure to emphasize the **Science** part of **Data Science**
- Spend more time **defining the questions** we are trying to answer to ensure we have the correct data to address the problem.
- Ensuring that sources of data are well known and robust enough to generalize to wide populations – if not limit the scope.
- Know the limitations of the method
  - Machine Learning works best when inputs and outputs are well defined and a clear metric for success is established!
- Machine learning systems designed to make decisions that are fully automated are **very risky**. The combination of **human intuition** and expertise alongside machine learning systems is a better approach.

Engineering of Machine Learning Algos versus Software Development





Made in Lucidchart

Source: https://towardsdatascience.com/stoend-to-end-data-science-life-cycle-6387523b5afc