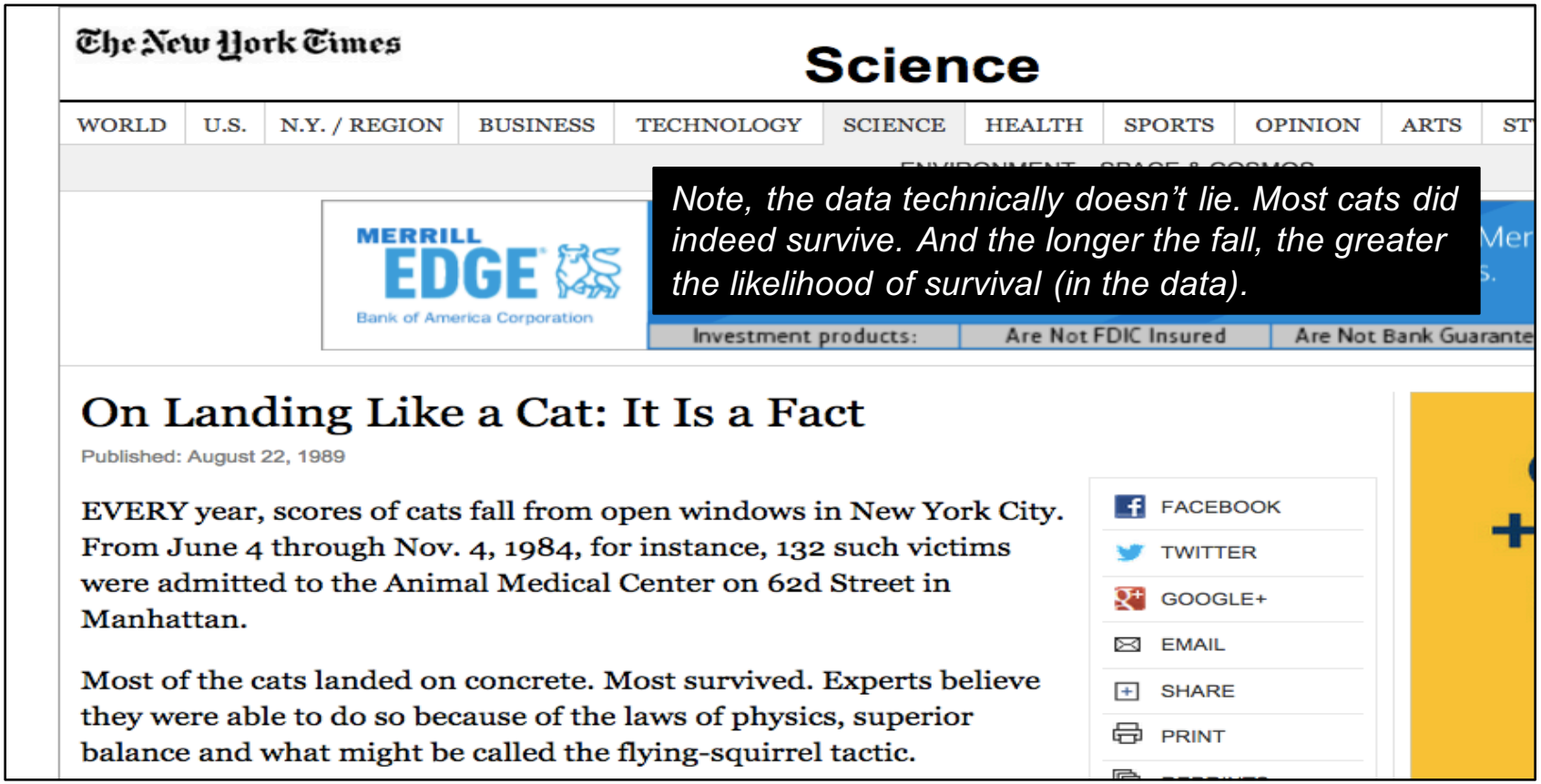
Introduction to Data Science

DISTRIBUTION BIASES:

SELECTION BIAS AND CONCEPT DRIFT

BRIAN D'ALESSANDRO

Fine Print: these slides are, and always will be a work in progress. The material presented herein is original, inspired, or borrowed from others' work. Where possible, attribution and acknowledgement will be made to content's original source. Do not distribute, except for as needed as a pedagogical tool in the subject of Data Science.



source: http://www.nytimes.com/1989/08/22/science/on-landing-like-a-cat-it-is-a-fact.html
https://www.youtube.com/watch?v=TGGGDpb04Yc

OCCAM'S RAZOR QUIZ

Conclusion derived from the data…

Even more surprising, the longer the fall, the greater the chance of survival.

Explanation 1 (per the article):

"Cats may be behaving like well-trained paratroopers," Dr. Jared Diamond, who teaches physiology at the University of California at Los Angeles Medical School, wrote in the August issue of the magazine Natural History.

Explanation 2 (per a reasonable data scientist):

Nobody brings their dead cat to the hospital, therefore this data suffers from selection bias. Cats that are clearly alive after a fall are more likely to be brought to the hospital, and more likely to survive. It is likely that the ambiguity of a cat's condition decreases with the height of the fall. Cat's are either obviously dead or obviously alive, which explains the trend in the data.

ANALYZING THE FALLING CAT ANALYSIS

It's always good to think through an analysis and sampling using the language and tools of probability.

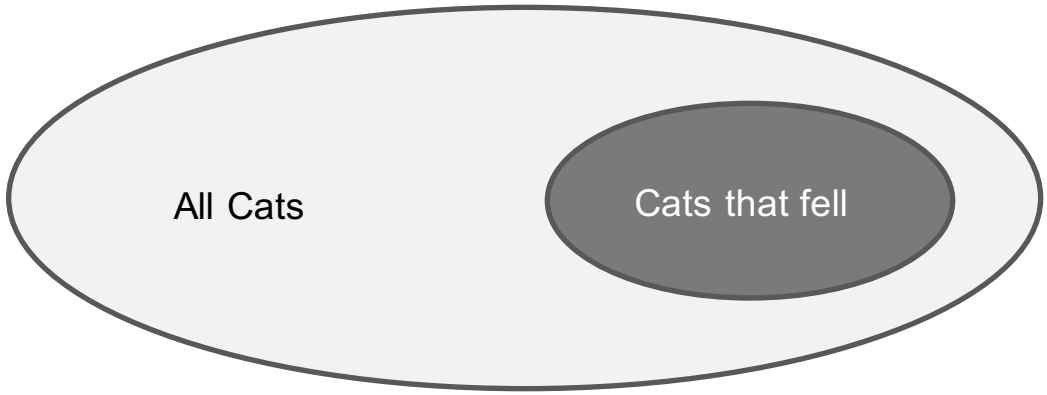
I.e.,

As a starting point, we're interested in cat survival. So let's say the we want to estimate using some data: *P(Survive)*. Note that specifying this probability essentially defines the problem, so being precise is incredibly important!

But is this precise enough? How might we change this to reflect the analysis? Should we modify event of interest ("Survive" vs something else)? Should we make it a conditional probability?

ANALYZING THE FALLING CAT ANALYSIS

Let's instead focus on *P(Survive 1 week)*. Let's also be more precise about the sub-population. We don't want all cats, we want all cats that fall from a window.



So let's reformulate the problem as estimating *P(Survive 1 Week | Fell)*.

ANALYZING THE FALLING CAT ANALYSIS

Can we estimate *P(Survive 1 Week | Fell)* from data collected by the vet?

Questions to ask:

1. Does the vet data represent all cats that fell?
2. If not, is the data missing at random?

MAR in this case is equivalent to saying:

P(Go to vet | Fell) = P(Go to vet| Fell, State of cat at Fall)

ANALYZING THE FALLING CAT ANALYSIS

We can use the law of total probability to work this out.

P(Survive 1 Week | Fell) =

*P(Survive 1 Week | Fell, ! Clearly Dead) *P(! Clearly Dead | Fell) +*

*P(Survive 1 Week | Fell, Clearly Dead) * P(Clearly Dead | Fell)*

*= P(Survive 1 Week | Fell, ! Clearly Dead) *P(! Clearly Dead | Fell)*

Let's assume the sampling mechanism is: *If ! Clearly Dead => Go to vet*

Thus:

P(Survive 1 Week | Fell, Sampled) = P(Survive 1 Week | Fell, ! Clearly Dead)

ANALYZING THE FALLING CAT ANALYSIS

So the reporting implies the following:

P(Survive 1 Week | Fell, Sampled) = P(Survive 1 Week | Fell)

But in reality they are actually measuring:

P(Survive 1 Week | Fell, Sampled) =

P(Survive 1 Week | Fell) / P(! Clearly Dead | Fell)

The degree of bias depends on how low *P(! Clearly Dead | Fell) is*.

SELECTION BIAS

Every analysis starts by drawing a data sample ***S*** from a population ***D.***

Each instance is characterized by a set of features ***(X,Y)***

If being in the sample ***S*** is independent of ***X*** and independent of ***Y***, the sample is unbiased:

i.e. ***P(S|X)=P(S) and P(S|Y)=P(S)***

Else the sample is biased

Recommended reading:
Zadrozny, Bianca. "Learning and evaluating classifiers under sample selection bias." *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004.

TWO TYPES OF SELECTION BIAS

Bias only depends on Target variable: ***P(S|X,Y) = P(S|Y)***

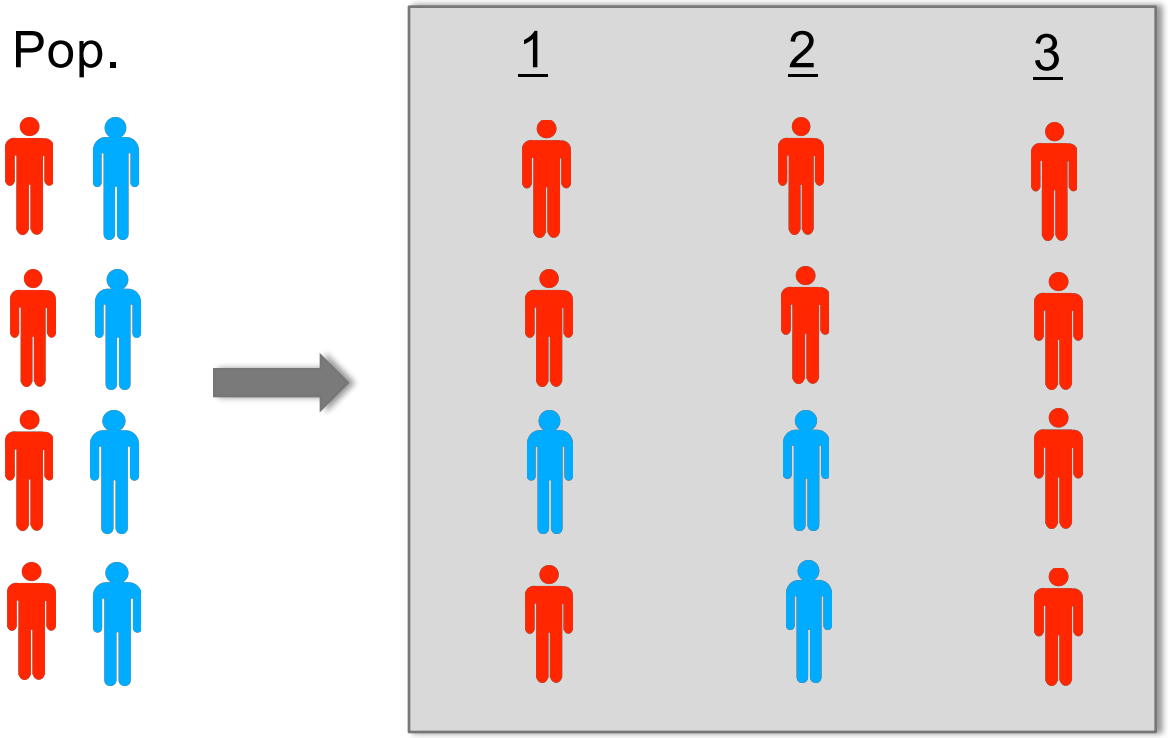
- This type of bias is common, intentional and often justified
- A bi-product of stratified sampling, up/down sampling
- Is usually done to improve learning
- Impacts prior probability (base rate) and is easily corrected
- Needs to be corrected when running evaluation

Bias only depends on feature vector: ***P(S|X,Y) = P(S|X)***

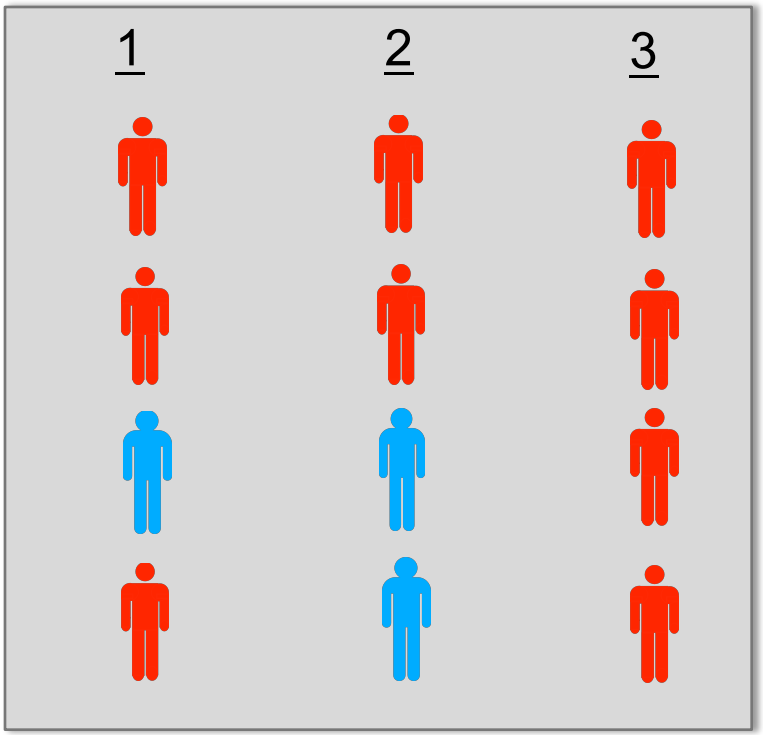
- This type of bias is common, but often unintentional
- A bi-product of business rules or constraints
- Can hurt learning and cause unintended consequences
- Can be corrected when certain conditions are true

Today's Focus

SELECTION BIAS – TOY EXAMPLES



SELECTION BIAS – TOY EXAMPLES



$P(S1) = 0.5$
$P(S1|R) = 0.75$
$P(S1|B) = 0.25$

$P(S2) = 0.5$
$P(S2|R) = 0.5$
$P(S2|B) = 0.5$

$P(S3) = 0.5$
$P(S3|R) = 1$
$P(S3|B) = 0$

SELECTION BIAS – IMPLICATIONS

Selection bias within data affects generalizability of results and potentially the identifiability of model parameters.

Generalizability:
Does your model represent the population at large?
Does your prediction match the production results?
Is your statistic representative of the greater population?

Identifiability:
Can you learn a model, parameter or statistic given the data at hand?

e.g,
in previous example, sample 3. Let's assume we want to know the average sales for blue figures, i.e., E[Sales|B]. Because P(Samp3|B)=0, we can not learn this parameter from the data.

SELECTION BIAS – EXAMPLE

Sometimes, selection bias is almost intentional, and is a rational decision caused by business and economic factors.

Credit Risk Modeling

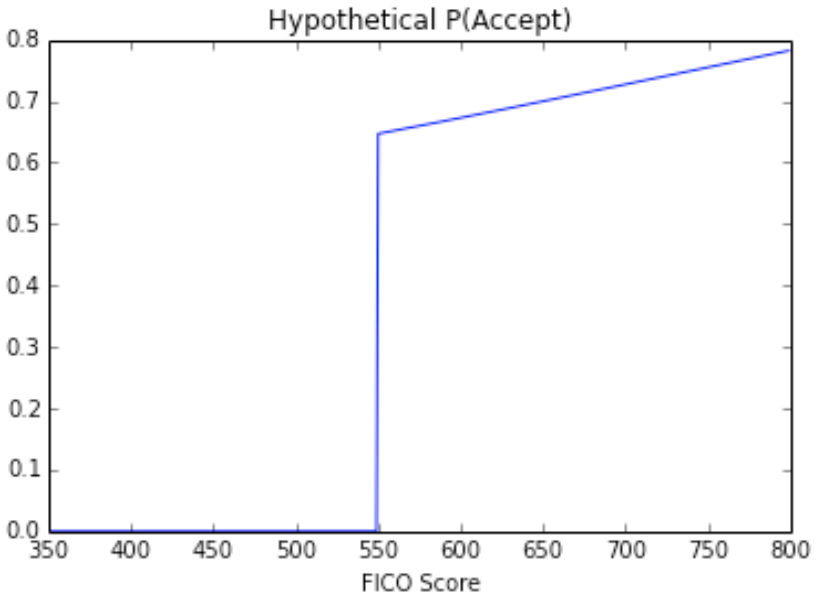
Goal: Predict P(Default in 6 mo's | Application Data)

Method:

1. Sample new credit card users, log app data
2. Observe 6 months, log if user defaults
3. Build a predictive model on sampled observations

Now what could go wrong?

SELECTION BIAS – EXAMPLE



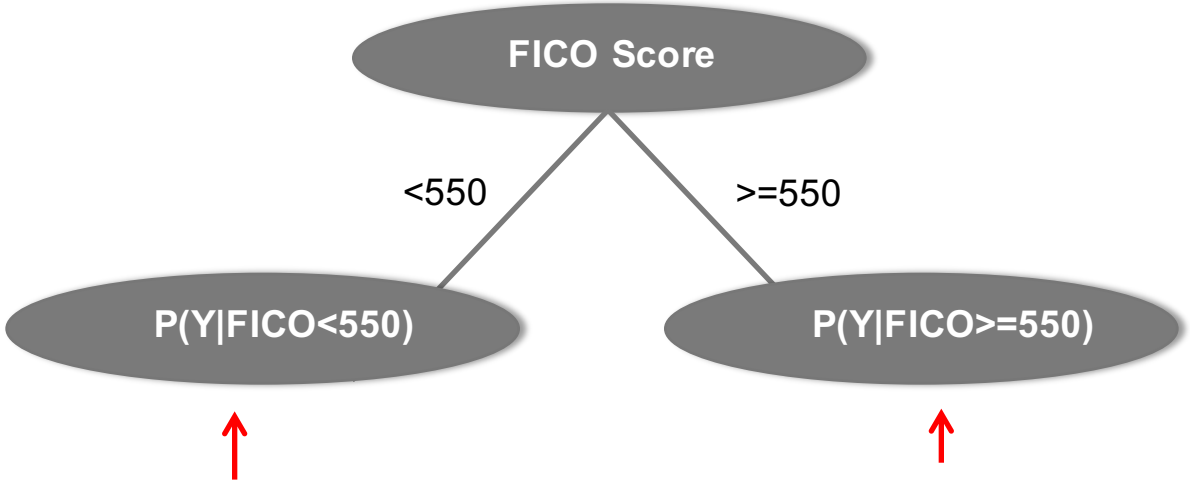
Q&A:

What is P(Sample|Fico<550)?

What implication does this have on future model building?

Is the deliberate selection bias worth it?

SELECTION BIAS – EXAMPLE



We have no data in this region, so it is technically not identifiable.

If the model is truly linear, we can extrapolate to this region, but that could be an expensive assumption!

We have the data to estimate this,…

but we can't generalize to the entire population.

SELECTION BIAS – WHAT TO DO

Selection bias is one of the biggest realities of production system data collection. What can you do about it?

1. Avoid It.

Design and use random sampling schemes as much as possible.

2. Adjust It.

In many cases you can statistically adjust for selection bias by weighting examples by 1/P(Samp|X) or Heckman Correction. In some cases your models will be fine (i.e., Logistic Regression w/ full data support).

3. Expect It.

Whether by design or accident, selection biases are likely to occur. Its always important to anticipate it and prepare for how it might affect your analysis.

AND ON TO CONCEPT DRIFT

aka. Non - Stationarity

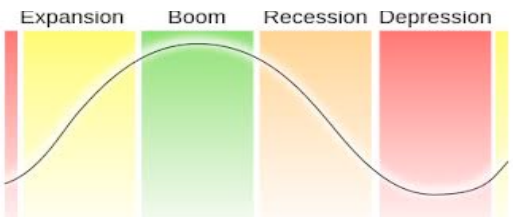
Defined simply as ***P(X), P(Y)*** or ***P(Y|X)*** that changes over time, and is almost a fact of life.

Example causes are…

Seasonality

Promotions (Exogenous Shocks)

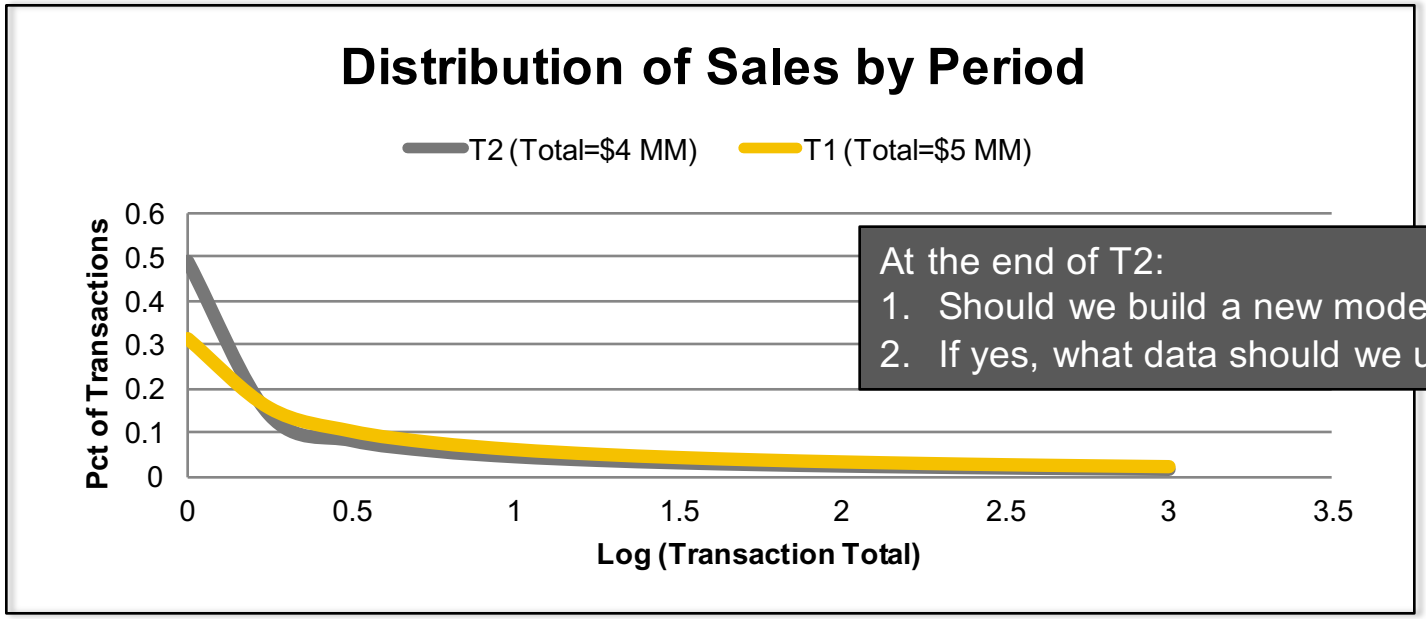
Economic Cycles



SIMPLE ILLUSTRATION

T1 (w/ TV Campaign)	T2 (No Campaign)	T3 (No Campaign)

We originally built a model here, to predict sales as a function of a user's history and demos.

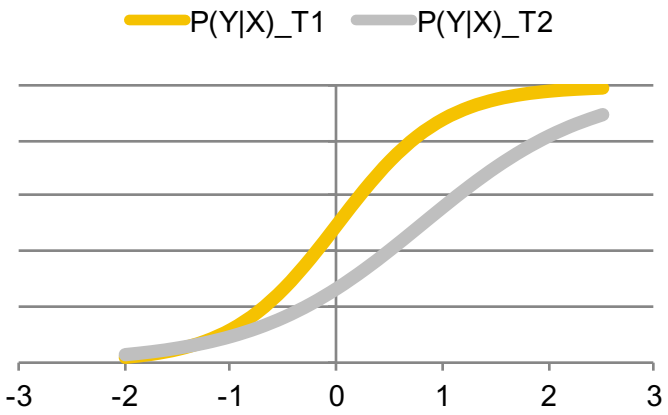


SIMPLE ILLUSTRATION

1. **Should we build a new model?**

We have proof that the distribution of transactions has changed, suggesting the underlying driver of purchase is different without the campaign. We know that T3 will look more like T2 than T1, so we probably should rebuild.

2. **If yes, what data should we use?**



We can pull both datasets and compare models. If models are very similar, pool data.

Otherwise, it's a balance between having more data and having the right data.

CONCEPT DRIFT TAKEAWAYS

Monitor predictive performance

You don't know future distributions, but model predictive performance should tell you when changes are happening.

Retrain as often as possible

The simplest way (in theory) to deal with an out-of-date model is to build a new one.

Test balance between data recency and data volume

This ties back to the classic bias-variance tradeoff, which is at the heart of many design decisions in data science.

Introduction to Data Science

DISTRIBUTION BIASES:

SELECTION BIAS AND CONCEPT DRIFT

BRIAN D'ALESSANDRO

Fine Print: these slides are, and always will be a work in progress. The material presented herein is original, inspired, or borrowed from others' work. Where possible, attribution and acknowledgement will be made to content's original source. Do not distribute, except for as needed as a pedagogical tool in the subject of Data Science.