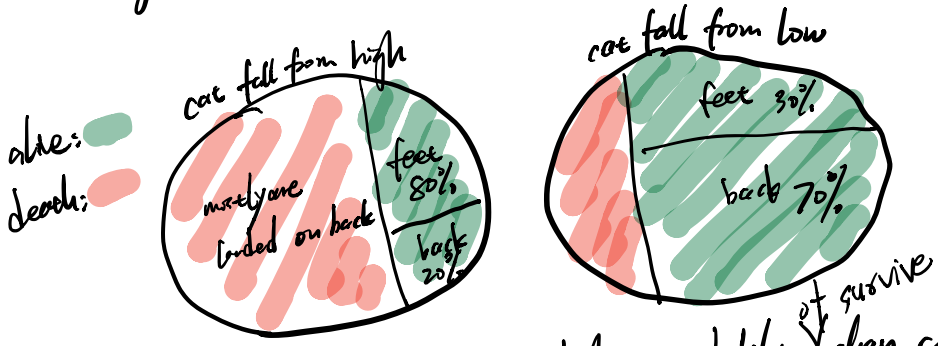
Cat Survival Experiment:

The longer the fall, the greater chance of survival.



cat land on feet has a higher probability ↓ than cat land on back even if they fall from higher place.

P(survive | high, feet) > P(survive | low, back)

Alive ones were sent to hospital.

- Among those cats fall from high, we say 70% of the cat land on feet survived so the $P(\text{survived} \mid \text{high}) = 70\% \times 80\% + 0\% \times 20\% = 56\%$ so 56% of cat fall from high are survived in the hospital.
- Among those cats fall from low, we say 80% of the cat land on feet survived, and 20% of cat land on back survived. so $P(\text{survived} \mid \text{low}) = 20\% \times 70\% + 80\% \times 30\% = 38\%$ in the hospital.
- The reason is because most cats fall from high are dead. If they are alive, they most likely land on feet. so they will have a high survival rate. However, most cats fall from low are not dead and sent to hospital. Lots of them land on back. so they will have a low survival rate after being sent to hospital. so the general probability

of survival of cats fall from low is lower than those fall from high.
However, the truth is most cats fall from high are clearly dead,
so they were not sent to hospital.

Reformulate the problem:

$P(\text{Survive 1 week} \mid \text{fell})$

Questions to ask:

1. Does the vet data represent all cat who fell ?
2. If not, is the data missing as random ? (MAR)

MAR in this case: $P(\text{go to vet} \mid \text{fell}) = P(\text{go to vet} \mid \text{fell, state of cat})$

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

$P(\text{Survive 1 week} \mid \text{fell}) = P(\text{Survive 1 week, ! clearly dead} \mid \text{Fell}) + P(\text{Survive 1 week, clearly dead} \mid \text{Fell})$

$P(A, B \mid C) = P(A \mid C)\, P(B \mid A, C)$

$P(\text{Survive 1 week, ! clearly dead} \mid \text{fell}) = P(\text{! clearly dead} \mid \text{fell}) \times P(\text{Survive} \mid \text{! clearly dead, fell})$

$P(\text{Survive 1 week, clearly dead} \mid \text{fell}) = P(\text{clearly dead} \mid \text{fell}) \times P(\text{Survive 1 week} \mid \text{clearly dead, fell})$

(first term: 0; second term: 0)

$\therefore P(\text{Survive 1 week} \mid \text{fell}) = P(\text{! clearly dead} \mid \text{fell}) \times P(\text{Survive 1 week} \mid \text{! clearly dead, fell})$

Equation (1)

Assume the sampling mechanism: If ! clearly dead → go to vet

thus.

$P(\text{Survive 1 week} \mid \text{fell. sampled}) = P(\text{Survive 1 week} \mid \text{fell. ! clearly dead})$

So the reporting implies: $P(\text{Survive 1 week} \mid \text{fell, sampled}) = P(\text{survive 1 week} \mid \text{fell})$

But in reality: $P(\text{survive 1 week} \mid \text{fell, sampled})$ Based on equation (1)

$= P(\text{survived 1 week} \mid \text{fell}) / P(\text{! clearly dead} \mid \text{fell})$

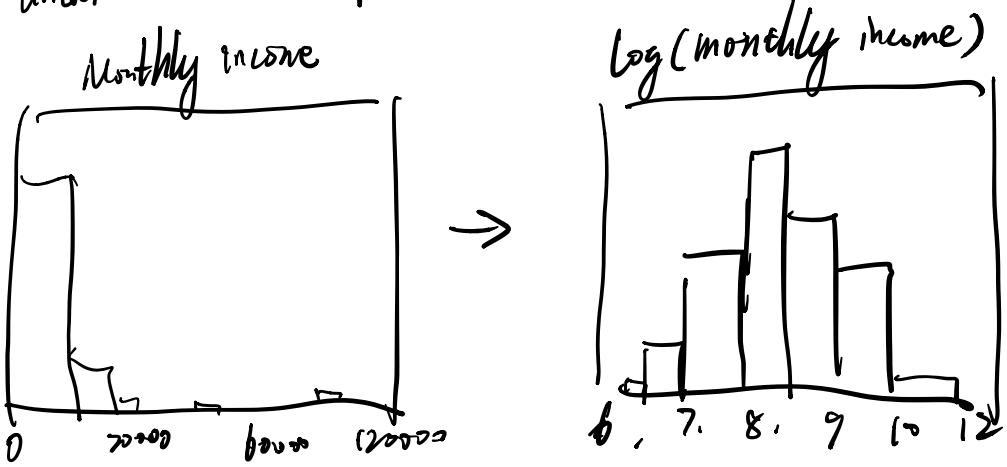
The degree of bias depends on how low $P(\text{! clearly dead} \mid \text{fell})$ is.

If fell from high, $P(\text{! clearly dead} \mid \text{fell})$ will be low. so the $P(\text{survive 1 week} \mid \text{fell. sampled})$ will be high. which explains the observation of the reporting

---

Knowing your data:

Understand its shape. The histogram is a great tool for this:



Outliers:

It's important to distinguish between an outlier from a normal process of the data-generating mechanism versus a outlier caused by some sort of failure within the system.

Missing data:

First to know whether the data is missing at random. look at the following page.

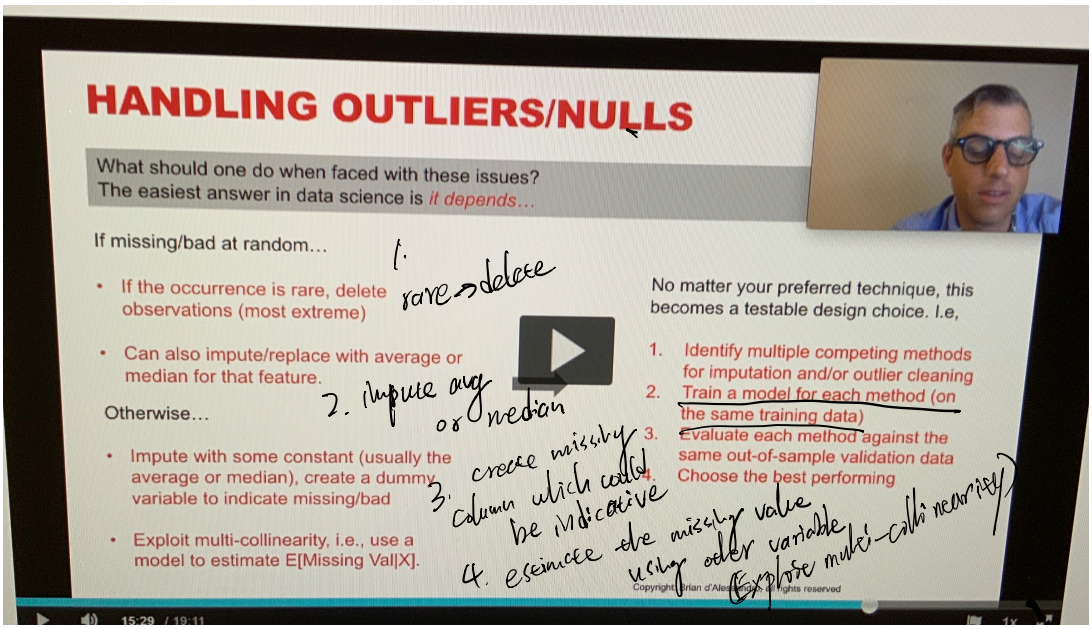
Why distributions are important when training ML model ?

Some ML models are designed to work best under distribution assumptions. Therefore. knowing with which distribution, we are working with. can help us to identify which models are best to use

Numerical data: Discrete (number of students) → PMF
continuous (heights and weights) → PDF

- PMF gives the probability that a value equal to certain value.
- PDF itself is not probability. they need to be integrated over the given range. to get the probability

Data Cleaning:



1. Identify multiple competing methods for inference
2. Train a model for each method using the same data
3. evaluate each method's performance against the same out-of-sample data
4. choose the best performing method

---

selection bias: the sample can not represent the general population.

selection bias affects the generalizability of results and potentially the identifiability of model parameters.

Generalizability:
Does your model represent the population of interest?
more loosely: average std. $y=f(x)$

ex: estimate the avg(height) of the world population, but you went to elementary school, so the avg(school population) can not represent the whole population.

Identifiability: Any estimate derived from the sample that do not contain our population of interest is not identifiable.

ex: you want to estimate the avg(height) of women, but your sample contains all men, so the avg(sample) cannot represent the population of interest (no women) ← this is our population of interest.

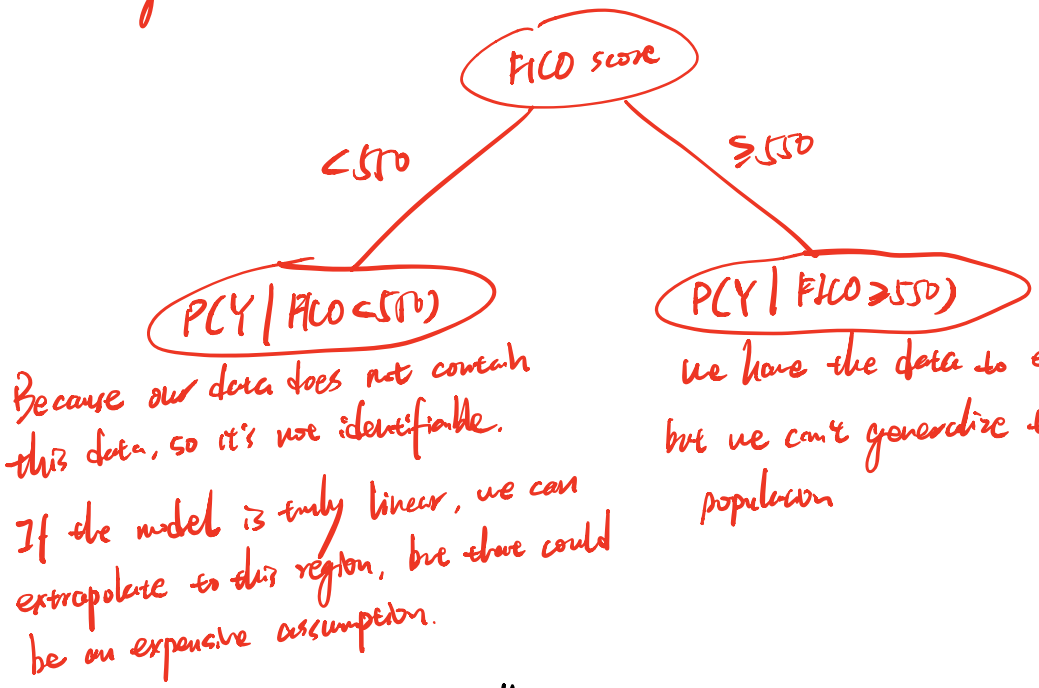
Credit Risk Modeling:

Goal: Predict P(Default in 6 mo's | application data)

Method:

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

In this case, since we don't approve the credit card for user with FICO < 550. we don't have the data for these users. so our sample only contains users with > 550 and have missing data.



Because our data does not contain this data, so it's not identifiable.

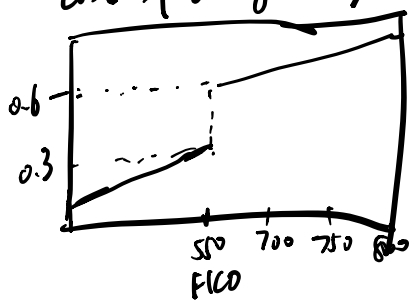
If the model is truly linear, we can extrapolate to this region, but there could be an expensive assumption.

we have the data to estimate this but we can't generalize to the entire population

Solution to the problem:

1. we can accept everybody. and the result will be ideal for learning. but it could be a huge investment, which could break the bank

2. take a hybrid approach. balance the need to learn and the cost of losing money.



could have high variance for the estimate of user (FICO < 550)

Concept drift (Non-stationary)

sample used for model may be different from the population over time.

Example causes: seasonality, economic cycles. holiday promotions

Should we build a new model?

If the distribution has changed we should probably rebuild.

What data to use?

having more doesn't mean having the right data.

we know T3 would more like T2

we can pull a small portion of T2 data as the evaluation set.

and build two models with and with T1 to compare the performance.

the rest of T2

1. Monitor predictive measure constantly.
2. Retrain as often as possible (if the model degrade)
3. Test balance between data recency and data volume.