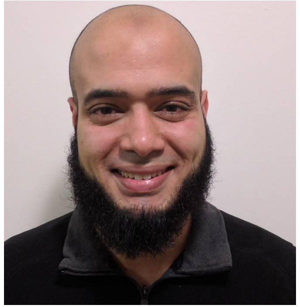
Machine Learning
Data Leakage

Mostafa S. Ibrahim
Teaching, Training and Coaching for more than a decade!

Artificial Intelligence & Computer Vision Researcher
PhD from Simon Fraser University - Canada
Bachelor / MSc from Cairo University - Egypt
Ex-(Software Engineer / ICPC World Finalist)



Data Leakage

- In English, leakage is the accidental escape of a fluid/gas through a hole
- So, in data leakage, data escapes from some test to the train set
- In other words, “*when the data you are using to train a machine learning algorithm happens to have the information you are trying to predict*.”
- A common consequence is having a **bit** higher val/test performance than the real performance
    - Depending on the leakage nature/size, this could be a minor or a major gap
    - With a major leakage, this might lead to overfitting
        - Test set performance is high, but the model doesn’t generalize!
    - In practice, typically a bit better performance (invisible leakage)

So Far

- We learned leakage through **pre-processing**
  - Split. Don’t touch test set. Do every statistics and learning based on train set only
  - Kaggle Competitions can do such leakage intentionally
- We learned leakage through **groups**
  - If there is something that generates many items, then we split on the key entities
    - Each video is split to 100 clips: split on the videos first
    - Each animal has 50 images. Split per animal. Then aggregate the pictures
      - If you augment/oversample an item, do that after you split first
      - Otherwise, the same example will be in train and test
- We learned leakage through **cross-validation**
- We learned leakage through **pipelines** [learned at work not internet]
  - First split data, then each module use this fixed split!

Where is the leakage?

EmployeeID	Title	ExperienceYears	MonthlySalaryGBP	AnnualIncomeUSD
315981	Data Scientist	3	5,000.00	78,895.44
4691	Data Scientist	4	5,500.00	86,784.98
23598	Data Scientist	5	6,200.00	97,830.35

src

Where is the leakage?

SubscriberID	Group	DailyVoiceUsage	DailySMSUsage	DailyDataUsage	Gender
24092091	M18-25	15.31	25	135.10	0
4092034091	F40-60	35.81	3	5.01	1
329815	F25-40	13.09	32	128.52	1
94721835	M25-40	18.52	21	259.34	0

src

Where is the leakage?

Education	Married	AnnualIncome	Purpose	LatePaymentReminders	IsBadLoan
1	Y	80k	Car Purchase	0	0
3	N	120k	Small Business	3	1
1	Y	85k	House Purchase	5	1
2	N	72k	Marriage	1	0

src

Where is the leakage?

- The patient visits the doctor with some symptoms. We have also his historical diagnosis and tests
- Given this information, we predict his current problem
- After the visit, we update his record with new symptoms/disease
- Assume the system **doesn’t record dates** for such events. Just **aggregate** all information together
- What is wrong if we trained a system on
  this aggregated data vs discovered diseases?

Where is the leakage?

- Our factory has machines and from time to time we maintain them and update the maintenance logs of the machine
- We would like to predict when a machine will fail
- You train on input: machine specs + maintenance history]
- What could go wrong?

Target leakage

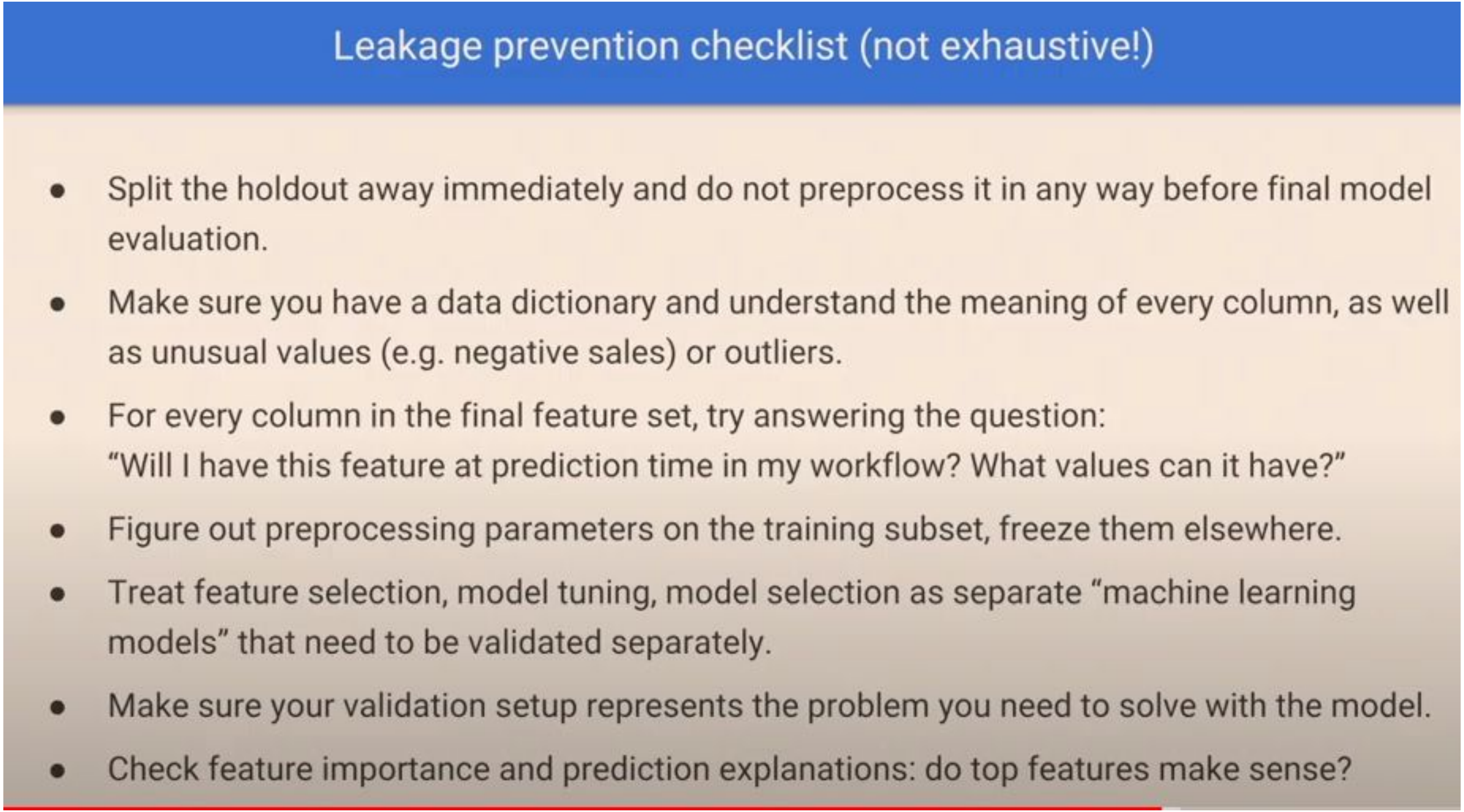
- Target leakage occurs when you include data in the model that would not be available **at the time of** prediction.
- For instance, if you are predicting **customer churn** and include features such as the number of customer service calls made, you may inadvertently include **future** information that wouldn't be known at the time of prediction
    - Be careful from a feature value **aggregated** over time!
- For instance, when predicting customer churn or retention, and data includes the number of days since the customer's last interaction with the company
    - This feature leaks information!
- Be careful from any feature computed based on future events (revene, future maintenance logs, future blood tests, etc)

Feature leakage

- Feature leakage happens when you include features that are **derived from the target variable**. For example, let's say you are predicting whether a loan will default or not, and one of the features is the loan status from the previous month. Including this feature would leak information about the target variable into the training data, leading to an overly optimistic model
- During EDA, analysts may introduce or change columns to better present/cluster the data. Some of these columns are based on the target columns
  - An an ML folk receiving such modified data, you don't know this history!
  - Try to understand the source of your data
- Carefully check features that are **highly correlated** with the target

Time-based data leakage

- This needs first understanding time-series data



[Src](#) - In future, review the video

"Acquire knowledge and impart it to the people."

"Seek knowledge from the Cradle to the Grave."