Hindsight Bias: How to deal with label leakage at scale

Till Bergmann (PhD), Senior Data Scientist
tbergmann@salesforce.com

24 Hours in the Life of Salesforce

B2C

Scale

1.5B
emails sent

4M
orders

260M
social posts

3B
Einstein predictions

888M
commerce page views

B2B

Scale

44M
reports and dashboards

41M
case interactions

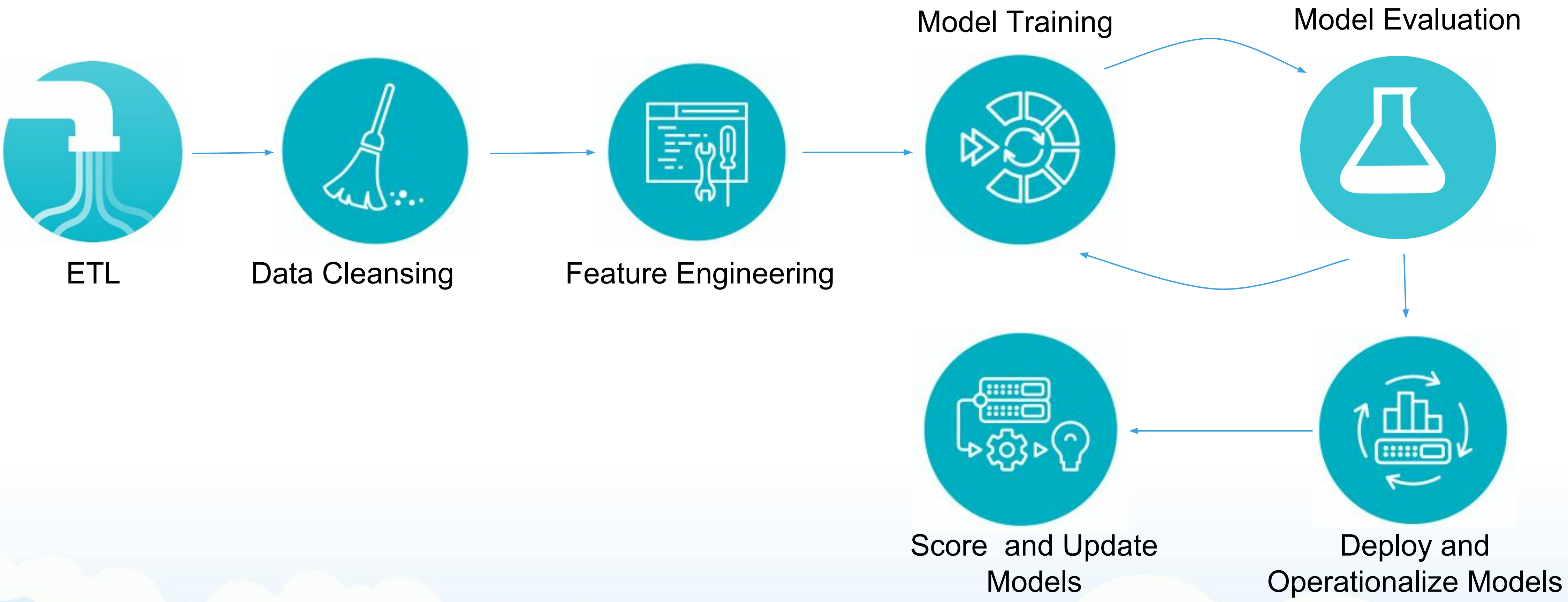
3M
opportunities created

2M
leads created

8M
cases logged

Source: Salesforce March 2018.

The typical Machine Learning pipeline



ETL

Data Cleansing

Feature Engineering

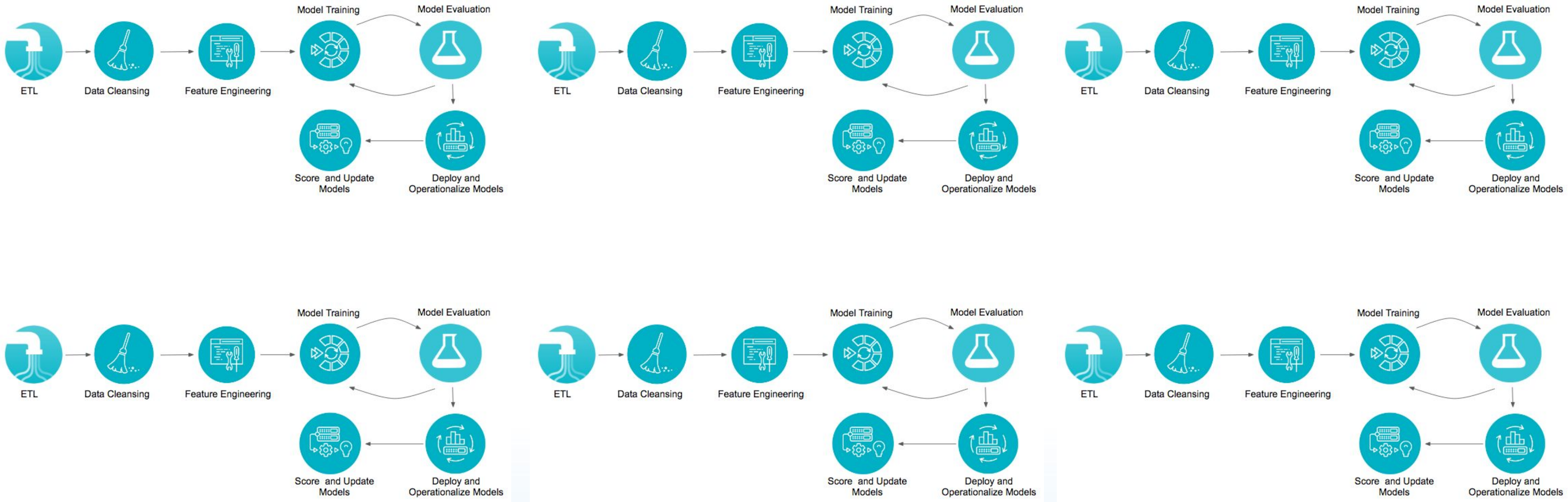
Model Training

Model Evaluation

Deploy and Operationalize Models

Score and Update Models

Multiply it by M*N (M = customers; N = use cases)



Problems with enterprise data

Not enough data scientists to hand tune each model

- We don’t know the specific business use case and data
- Each step in the pipeline needs to be automated

Messy data

- Nobody likes data entry - missing fields, typos
- Automated business practices can lead to patterns in the data
- Custom fields get added, removed or deprecated at any time

No historical data

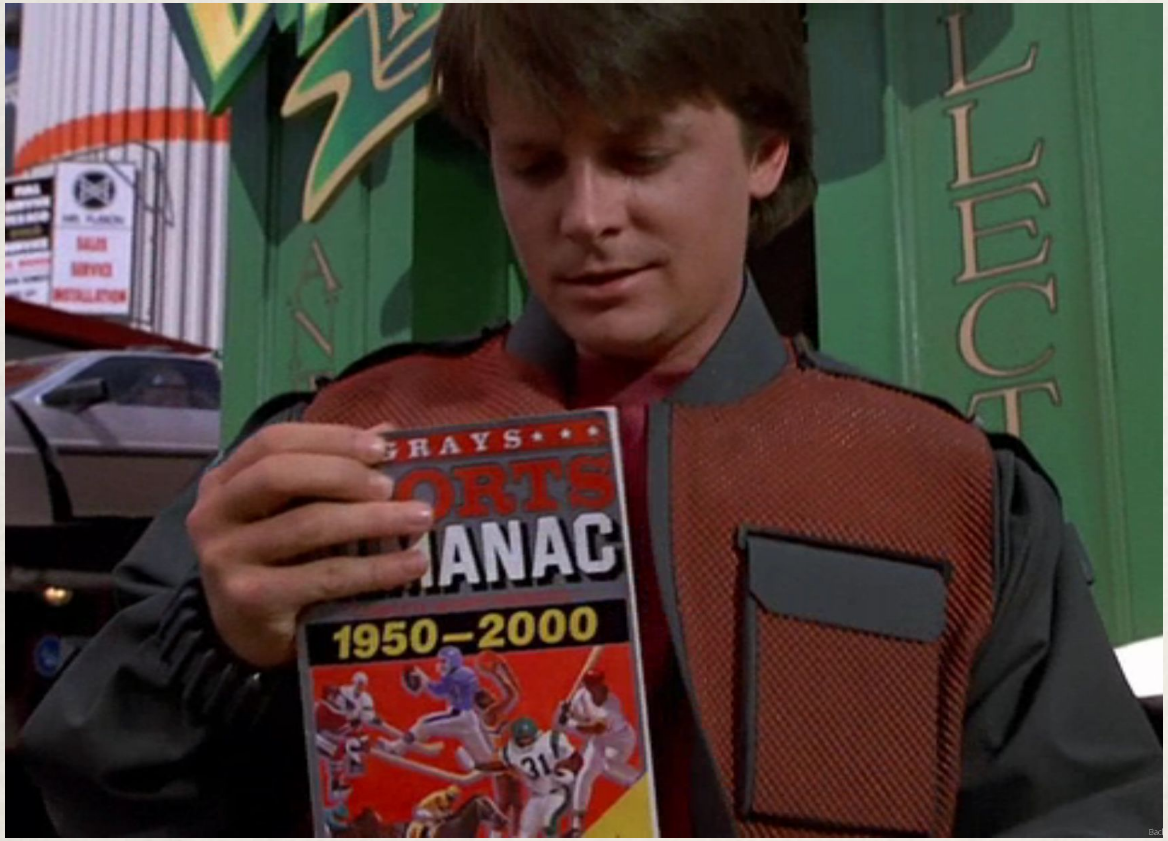
- Impossible to keep track of value changes in every field
- Cold start problem

What is hindsight bias?

Label/data leakage

Back to the future

Knowing things you shouldn't know



The Weather

Today – Mostly sunny, high near 80 low near 60. The chance of precipitation is near zero through tonight. Thursday – Variable cloudiness with a high near 80. Yesterday's temperature range: 80-64. Details Page C3.

Hill Valley Telegraph

Index

4 Sections
72 Pages

Amusements	B 6	Obituaries	C 4
Classified	C 6	Sports	D 1
Comics	B 5	Style	B 1
Editorials	A20	Radio	B11
Financial	D10	Movies	C 1

Vol. XVII, No. 32 | COMPLETE NEWS SERVICE A. B. S. NEWS PHOTOS | PUBLISHED DAILY | THE NEWSPAPER THE PEOPLE DEPEND UPON | PRICE: 10 CENTS

BIFF WINS AGAIN

Khruschev Offers Dates for Summit

An immediate investigation is assured and indications are that some new light will be shed on the situation in the near future. Available facts seem vague but authorities feel that time will disclose some means of arriving at a solution.

Future plans will, of necessity, have great bearing on the situation as it now stands. Decisions will have to be made of the actual planning of the project will take considerable time but it is felt that these steps are very important.

Future plans will, of necessity, have great bearing on the situation as it now stands. Decisions will have to be made of the actual planning of the project will take considerable time but it is felt that these

Of no less importance was the common recognition shown of the fact that any menace from without to the peace of our continents concerns all of us and therefore properly is a subject for consultation and cooperation. This was reflected in the instruments adopted by the conference.

An immediate investigation is assured and indications are that some new light will be shed on the situation in the near future. Available facts seem vague but authorities feel that time will disclose some means of arriving at a solution.

A suggestion that public hearings on applications be limited to one every six months was taken under ad-

PAUL CRUMRINE
Staff Writer

While the final outcome of this situation has yet to be determined, it is possible that when all the facts come together at the conference now scheduled for next week, that a decision will be made to possibly satisfy the needs and demands of all parties.

Of no less importance was the common recognition shown of the fact that any menace from without to the peace of our continents concerns all of us and therefore properly is a subject for consultation and cooperation. This was reflected in the instruments adopted by the conference.

A suggestion that public hearings on applications be limited to one every six months was taken under advisement by the commission.

Future plans will, of necessity, have great bearing on the situation as it now stands. Decisions will have to be made of the actual planning of the

It would appear that the preliminary inquiry into this matter has in fact not settled any of the minor differences arising from the situation but rather has aggravated the mood of those petitioning for more local involvement by the council.

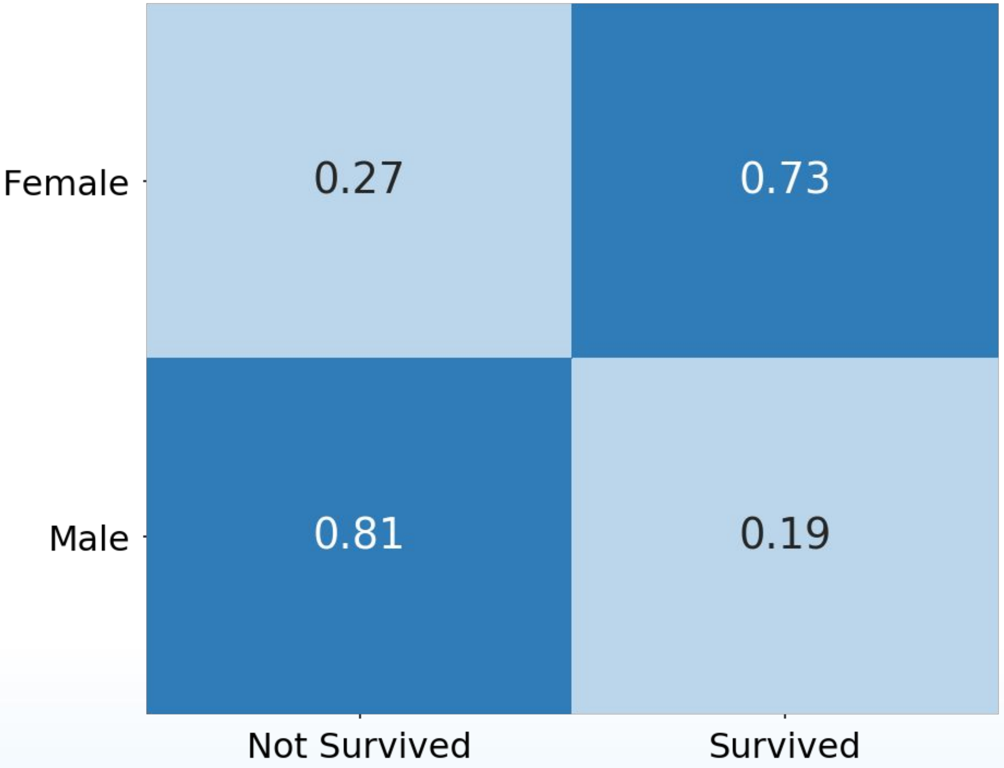
The facts regarding the situation remain the same, state the authorities. Details concerning the action have been given a preliminary investigation but it is felt that only by a more detailed study will the true facts become known.

The Mayor, meanwhile, has diplomatically kept a low profile, at least in public. Sources at City Hall confirm that until the council has finished, its private meetings concerning the issue, that the Mayor will have no public statement to make on the matter.

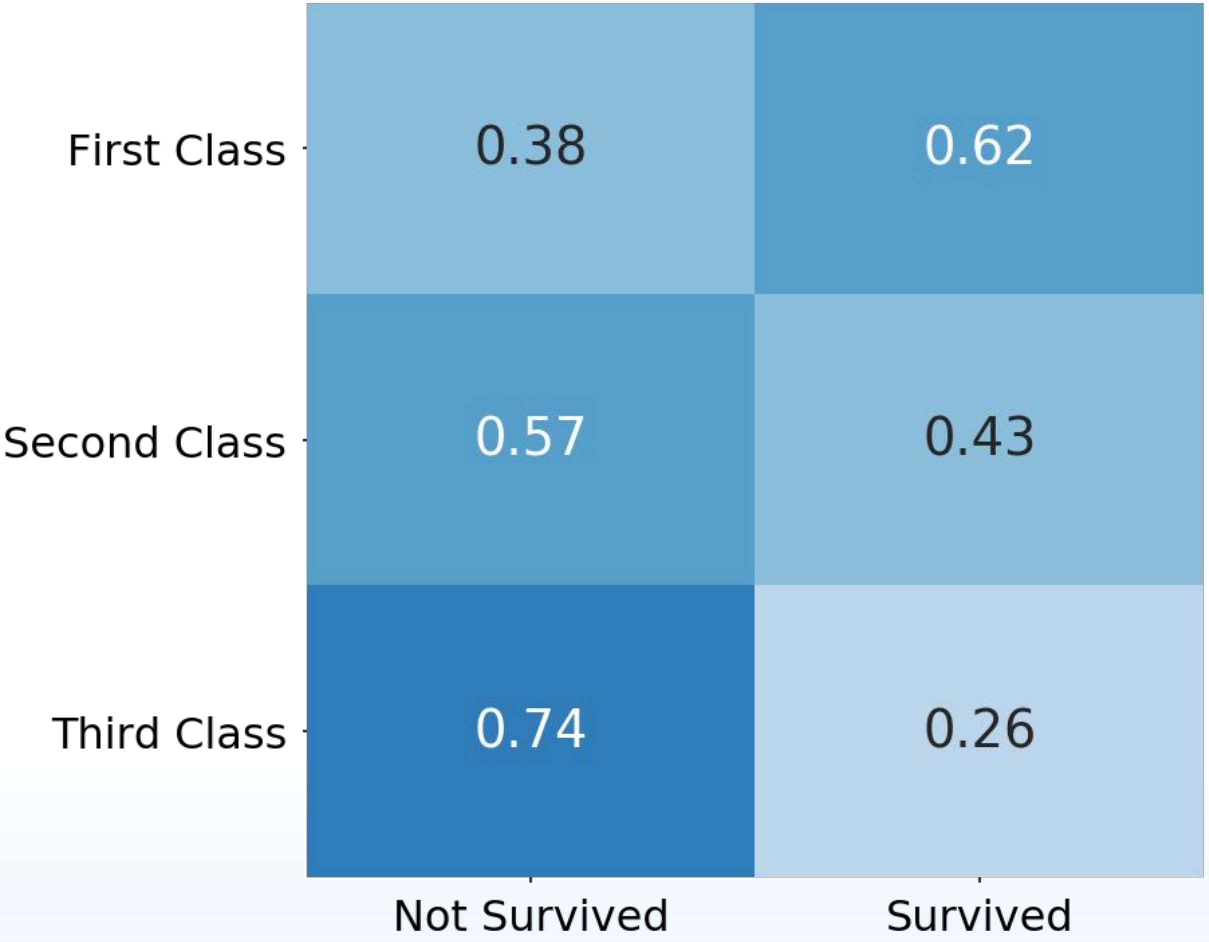
In any event, arrangements are going forward toward what is hoped will be a friendly meeting of both sides at the

A classic example

Predicting survival on the Titanic



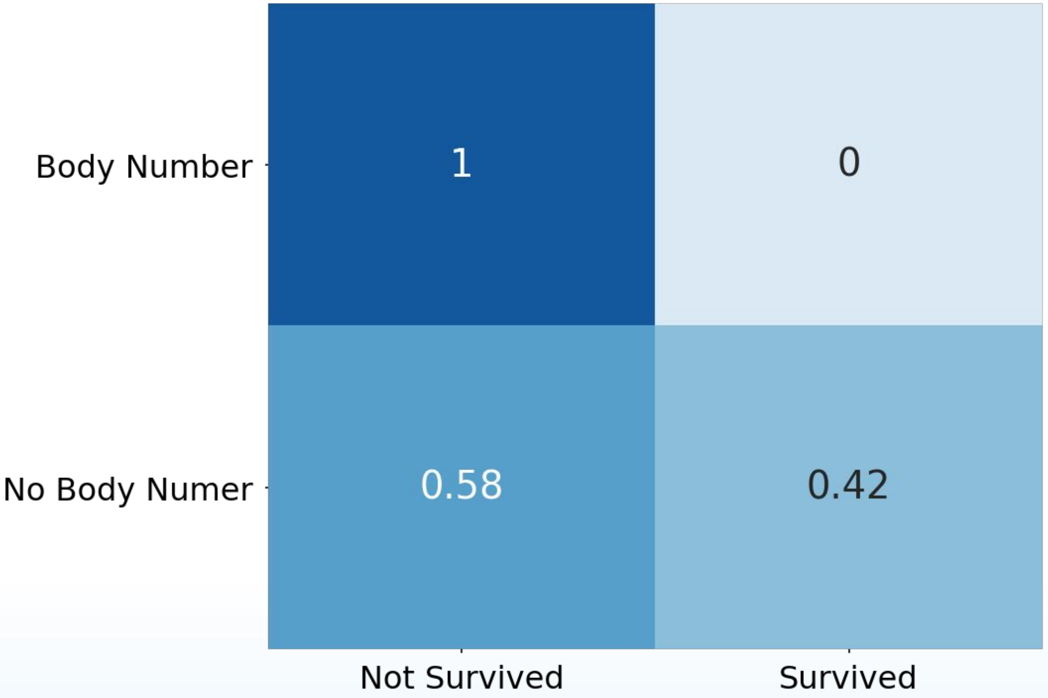
	Not Survived	Survived
Female	0.27	0.73
Male	0.81	0.19



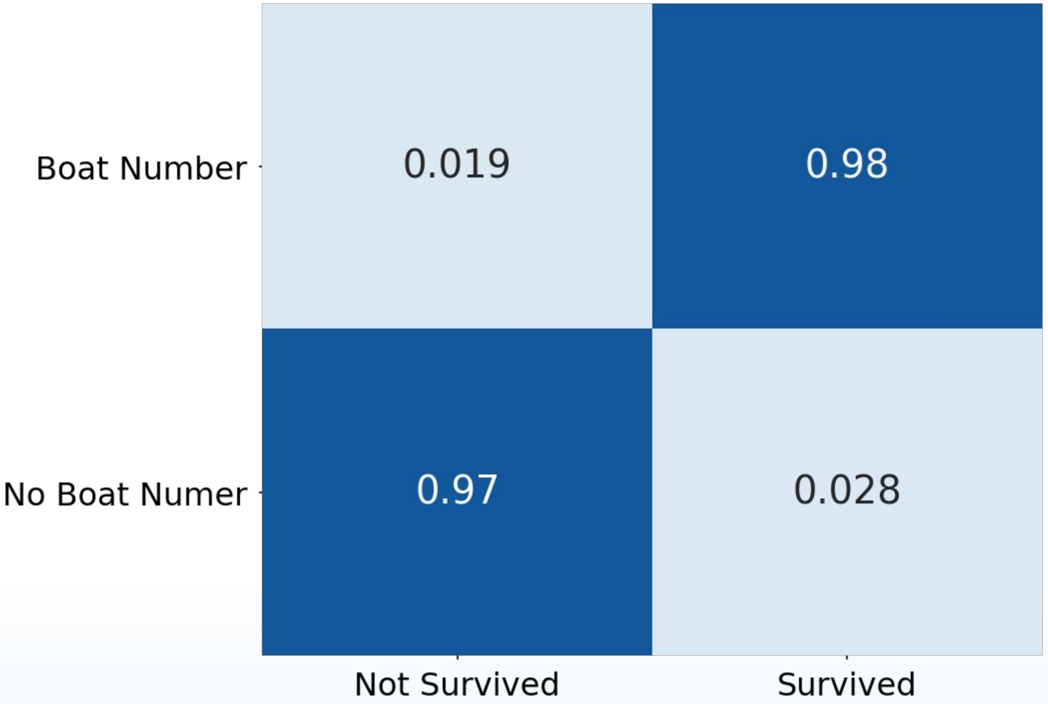
	Not Survived	Survived
First Class	0.38	0.62
Second Class	0.57	0.43
Third Class	0.74	0.26

A classic example

Predicting survival on the Titanic



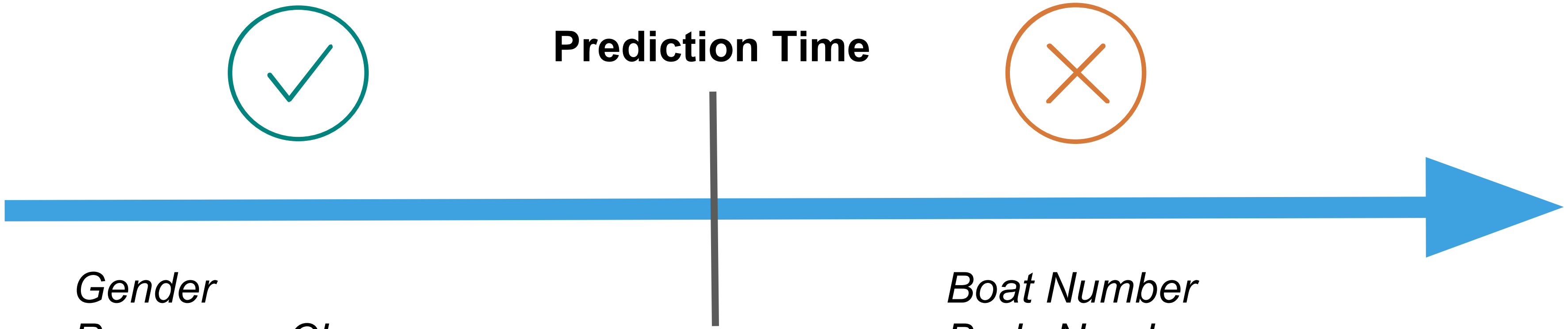
	Not Survived	Survived
Body Number	1	0
No Body Numer	0.58	0.42



	Not Survived	Survived
Boat Number	0.019	0.98
No Boat Numer	0.97	0.028

A classic example

Predicting survival on the Titanic



Prediction Time

✓

Gender
Passenger Class
...

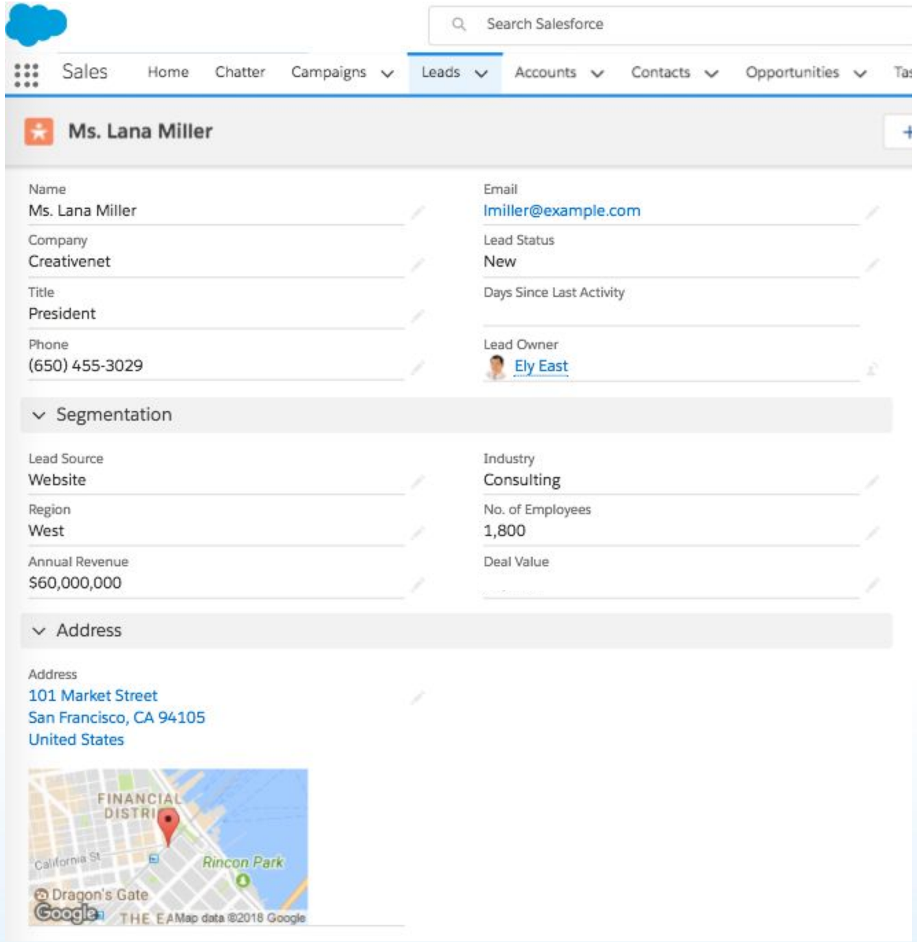
✗

Boat Number
Body Number

A modern example

Predicting lead conversion in Salesforce

Before Conversion



Ms. Lana Miller

Name
Ms. Lana Miller

Email
lmiller@example.com

Company
Creativenet

Lead Status
New

Title
President

Days Since Last Activity

Phone
(650) 455-3029

Lead Owner
Ely East

Segmentation

Lead Source
Website

Industry
Consulting

Region
West

No. of Employees
1,800

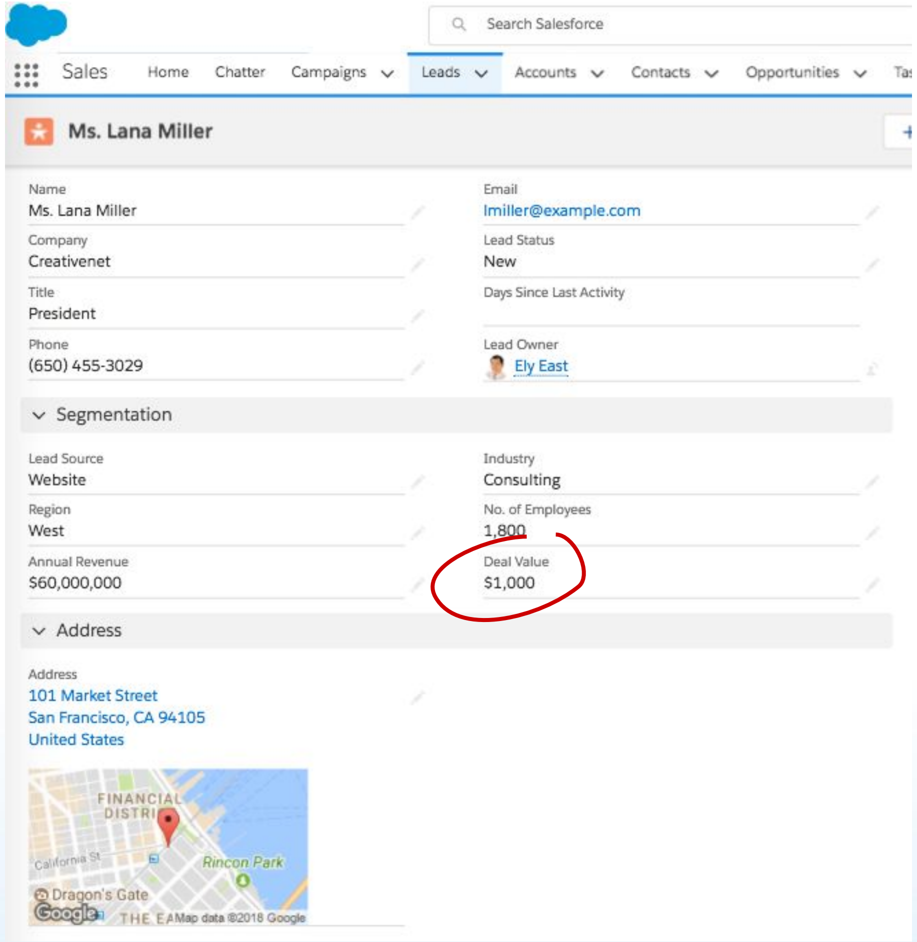
Annual Revenue
$60,000,000

Deal Value

Address

Address
101 Market Street
San Francisco, CA 94105
United States

After Conversion



Ms. Lana Miller

Name
Ms. Lana Miller

Email
lmiller@example.com

Company
Creativenet

Lead Status
New

Title
President

Days Since Last Activity

Phone
(650) 455-3029

Lead Owner
Ely East

Segmentation

Lead Source
Website

Industry
Consulting

Region
West

No. of Employees
1,800

Annual Revenue
$60,000,000

Deal Value
$1,000

Address

Address
101 Market Street
San Francisco, CA 94105
United States

Why does it even matter?

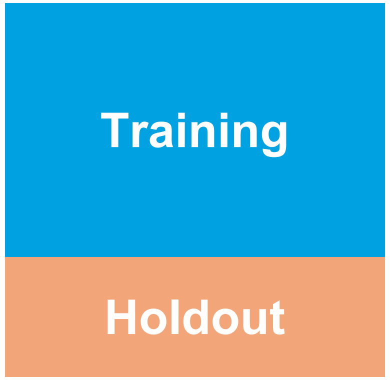
Good for betting, but not machine learning

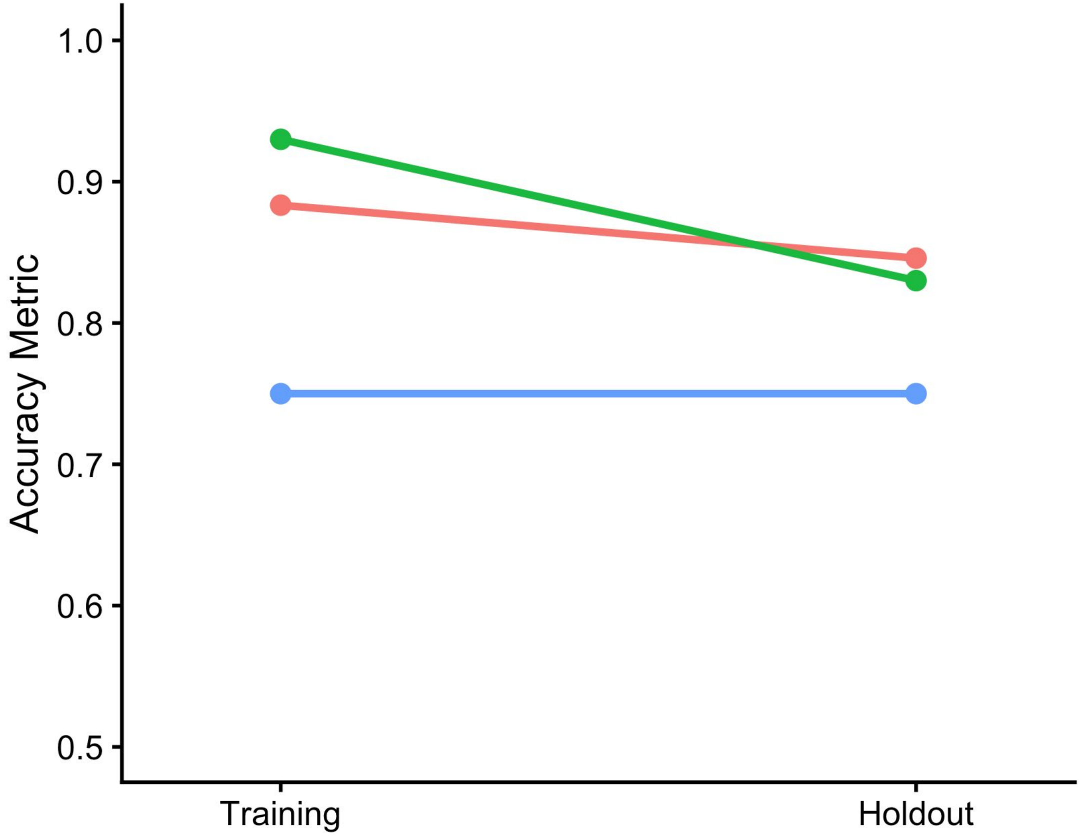
Effect on model performance

Traditional evaluation

Model relies on information not available at scoring time

- Model performance decreases for actual prediction
- Traditional evaluation pipeline is not sufficient



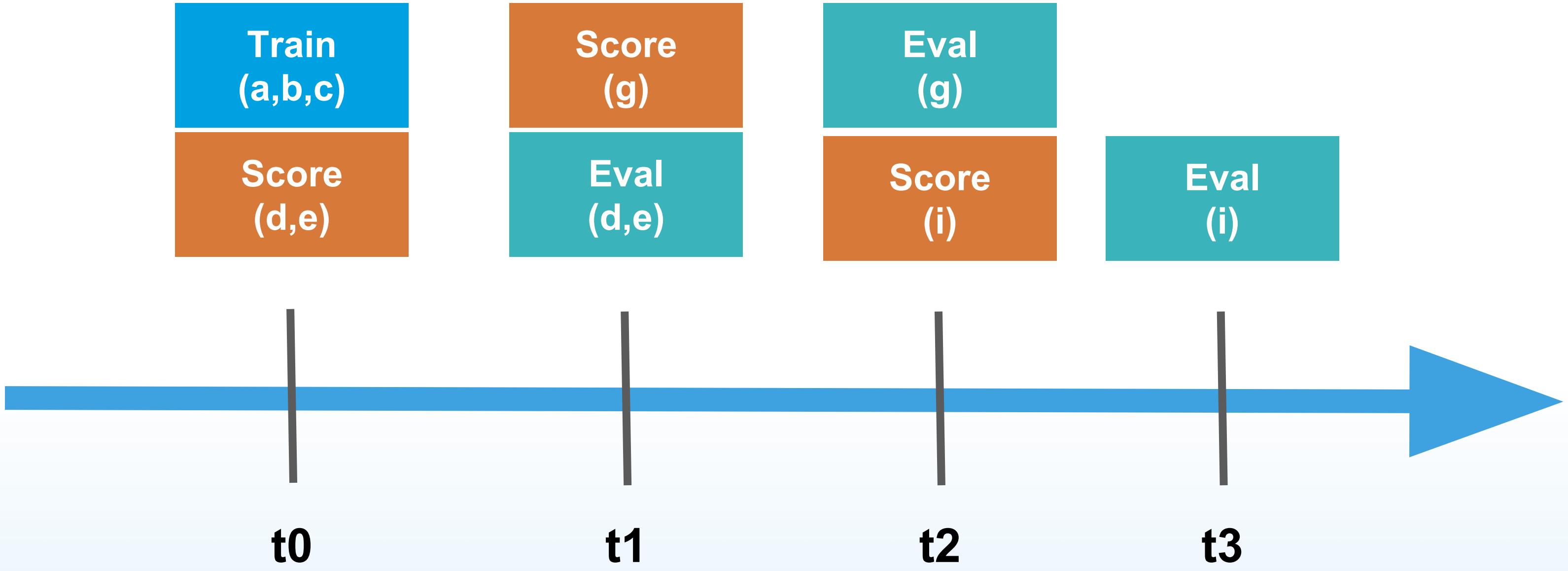


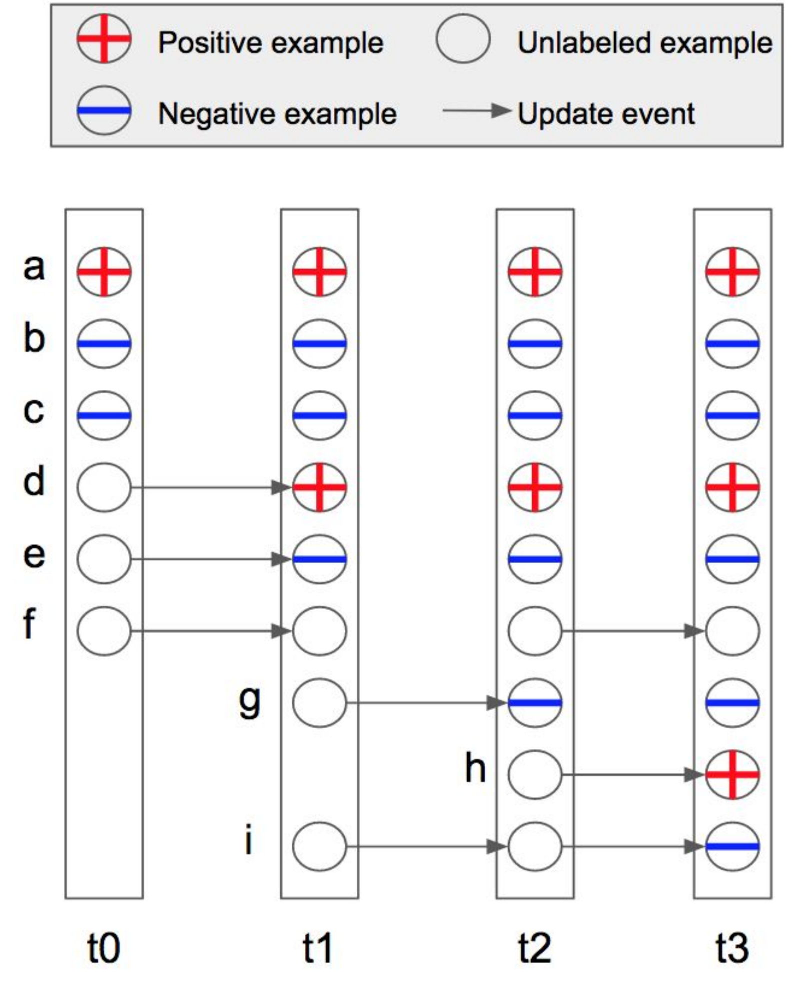
Effect on model performance

Time-based evaluation

Need to treat each record separately

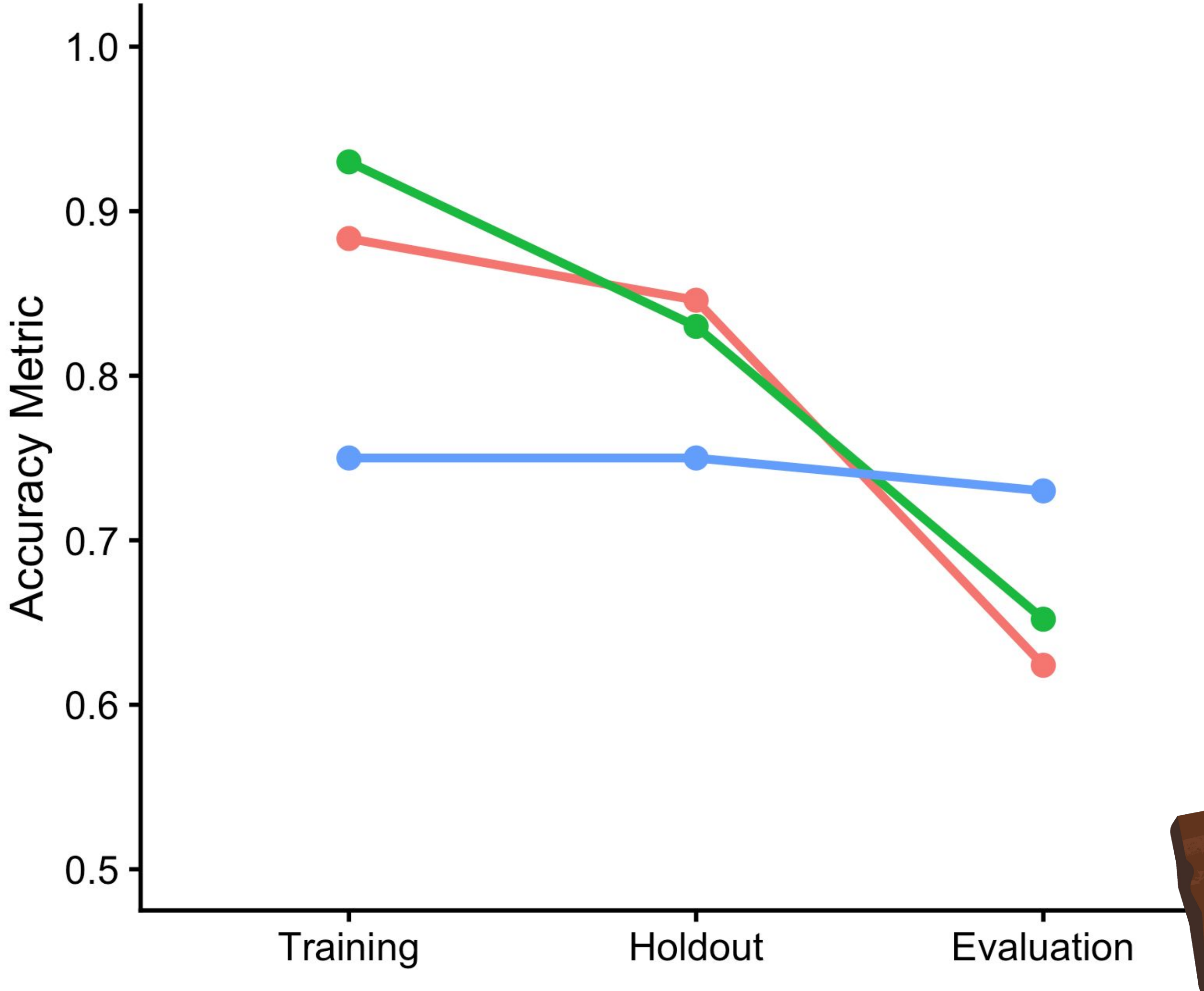
- Score and evaluate at different times





Effect on model performance

Time based evaluation



No leakers in evaluation data!

How do we solve it?

Solutions, and more problems

A simple start

What are some problems with this data?

Id	Name	Address	Phone	ClosedBy	ReasonLost	Amount	**Converted**	...
342	...	...	...	32212	-	$41k	True	
221	...	...	...	-	-	-	False	
098	...	...	...	86721	Unknown	-	False	
462	...	...	...	32212	-	$23k	True	
140	...	...	...	-	Competitor	-	False	

A simple start

What are some problems with this data?

- *ReasonLost* filled out means no conversion

Id	Name	Address	Phone	ClosedBy	ReasonLost	Amount	**Converted**	...
342	...	...	...	32212	-	$41k	True	
221	...	...	...	-	-	-	False	
098	...	...	...	86721	Unknown	-	False	
462	...	...	...	32212	-	$23k	True	
140	...	...	...	-	Competitor	-	False	

A simple start

What are some problems with this data?

- *ReasonLost* filled out means no conversion
- *Amount* filled out means conversion

Id	Name	Address	Phone	ClosedBy	ReasonLost	Amount	**Converted**	...
342	...	...	...	32212	-	$41k	True	
221	...	...	...	-	-	-	False	
098	...	...	...	86721	Unknown	-	False	
462	...	...	...	32212	-	$23k	True	
140	...	...	...	-	Competitor	-	False	

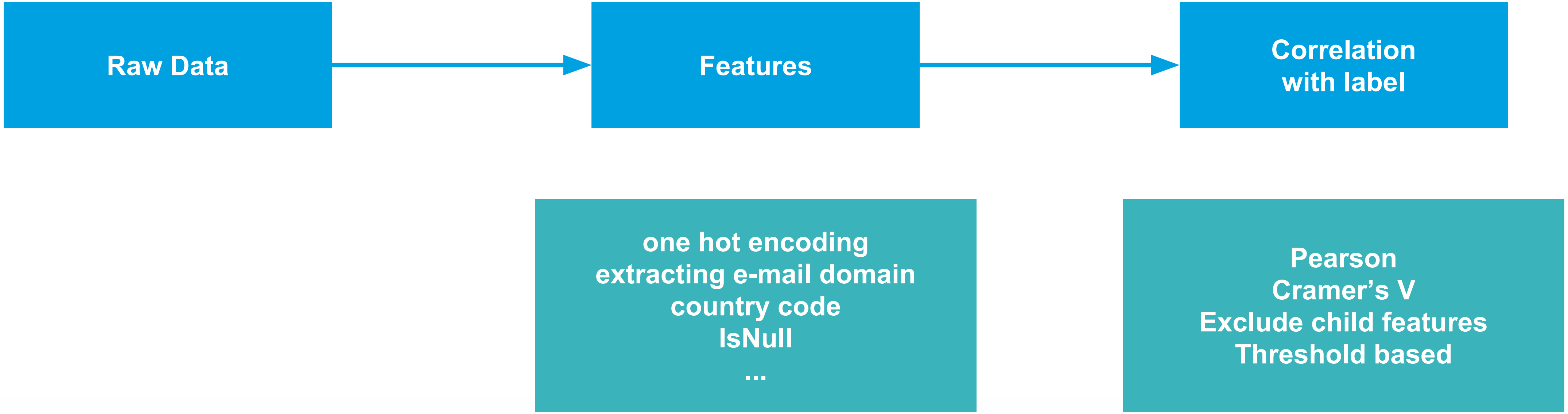
A simple start

What are some problems with this data?

- *ReasonLost* filled out means no conversion
- *Amount* filled out means conversion
- *ClosedBy* filled out, more likely to have conversion

Id	Name	Address	Phone	ClosedBy	ReasonLost	Amount	**Converted**	...
342	...	...	...	32212	-	$41k	True	
221	...	...	...	-	-	-	False	
098	...	...	...	86721	Unknown	-	False	
462	...	...	...	32212	-	$23k	True	
140	...	...	...	-	Competitor	-	False	

Catching features that are *too* good



Raw Data

Features

Correlation with label

one hot encoding
extracting e-mail domain
country code
IsNull
...

Pearson
Cramer’s V
Exclude child features
Threshold based

Does not solve everything

Data behaves in mysterious ways

Id	Name	Address	Phone	Expected Revenue	**Converted**	...
342	...	...	...	0		
221	...	...	...	0	False	
098	...	...	...	0		
462	...	...	...	15,000	True	
140	...	...	...	12,000	True	

Does not solve everything

Data behaves in mysterious ways

- Default value is not always *null*

Id	Name	Address	Phone	Expected Revenue	**Converted**	...
342	...	...	...	0		
221	...	...	...	0	False	
098	...	...	...	0		
462	...	...	...	15,000	True	
140	...	...	...	12,000	True	

Does not solve everything

Data behaves in mysterious ways

- Default value is not always *null*
- A value > 0 indicates conversion

Id	Name	Address	Phone	Expected Revenue	**Converted**	...
342	...	...	...	0		
221	...	...	...	0	False	
098	...	...	...	0		
462	...	...	...	15,000	True	
140	...	...	...	12,000	True	

Does not solve everything

Data behaves in mysterious ways

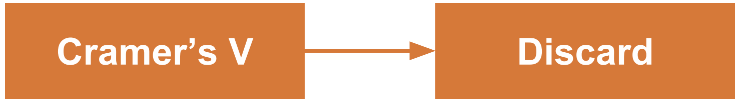
- Default value is not always *null*
- A value > 0 indicates conversion
- Auto-bucketizing can catch these cases

Id	Name	Address	Phone	Expected Revenue	**Converted**	...
342	...	...	...	0		
221	...	...	...	0	False	
098	...	...	...	0		
462	...	...	...	15,000	True	
140	...	...	...	12,000	True	

Does not solve everything

Data behaves in mysterious ways

- Default value is not always *null*
- A value > 0 indicates conversion
- Auto-bucketizing through decision tree can catch these cases



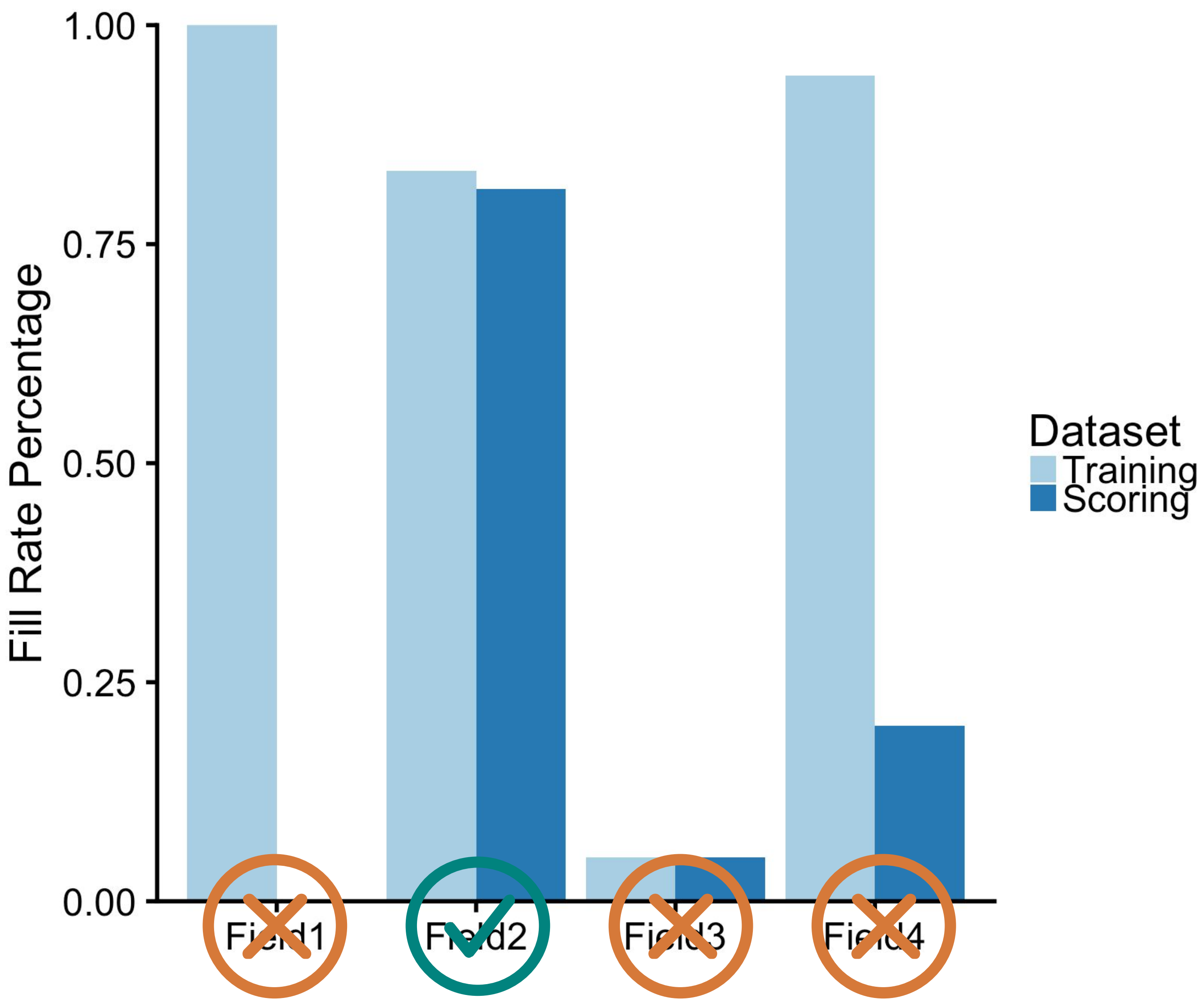
Cramer's V → **Discard**

Id	Name	Address	Phone	Expected Revenue	Bucketized	**Converted**	...
342	...	...	...	0	[1, 0, 0]		
221	...	...	...	0	[1, 0, 0]	False	
098	...	...	...	0	[1, 0, 0]		
462	...	...	...	15,000	[0, 1, 0]	True	
140	...	...	...	12,000	[0, 1, 0]	True	

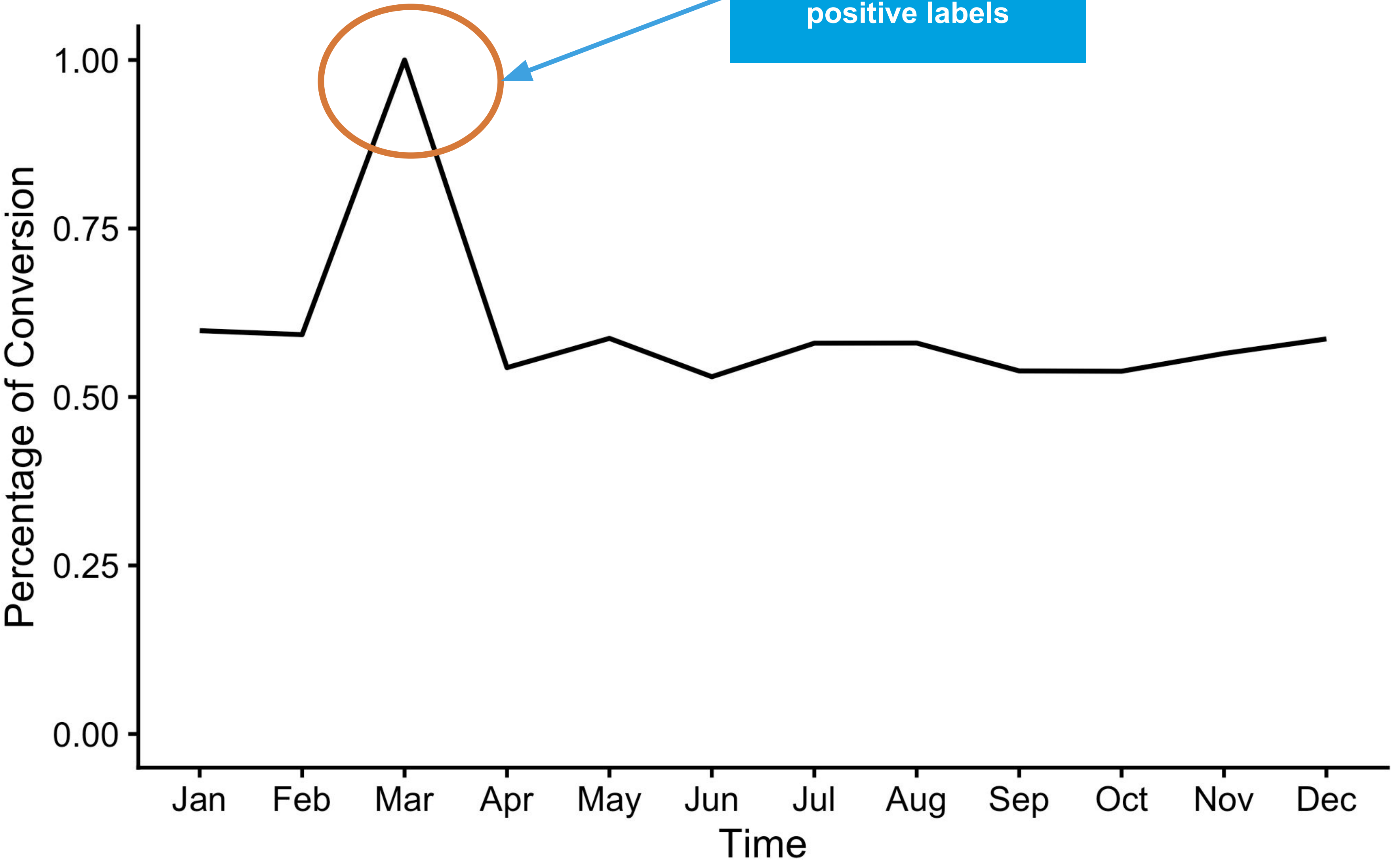
Change over time

So far, we have only talked about data at the same point in time

- But training and scoring data are rarely produced at the same time
- Training data is historical, scoring data is more current



Bulk uploads



Biased towards positive labels

Criteria to exclude

Low overall fill ratio

- No point in keeping a feature that is mostly null

Big discrepancy between training and scoring

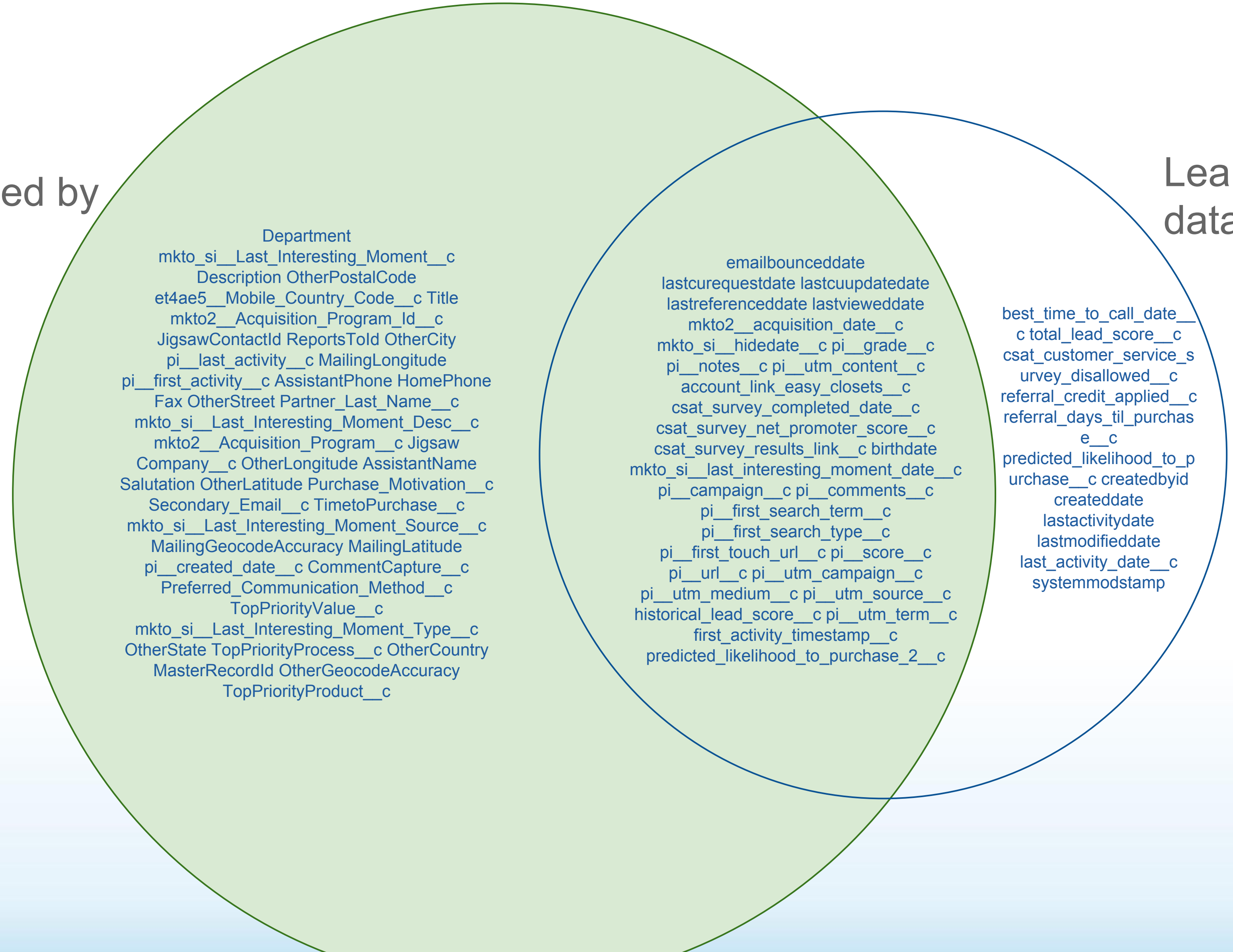
- Convert to probability distribution and compare with Jensen-Shannon Divergence

Skewed dates and ratios

- Be careful about including date features that might be inherently biased

AutoML vs Hand Tuning

Leakers removed by
AutoML: 73



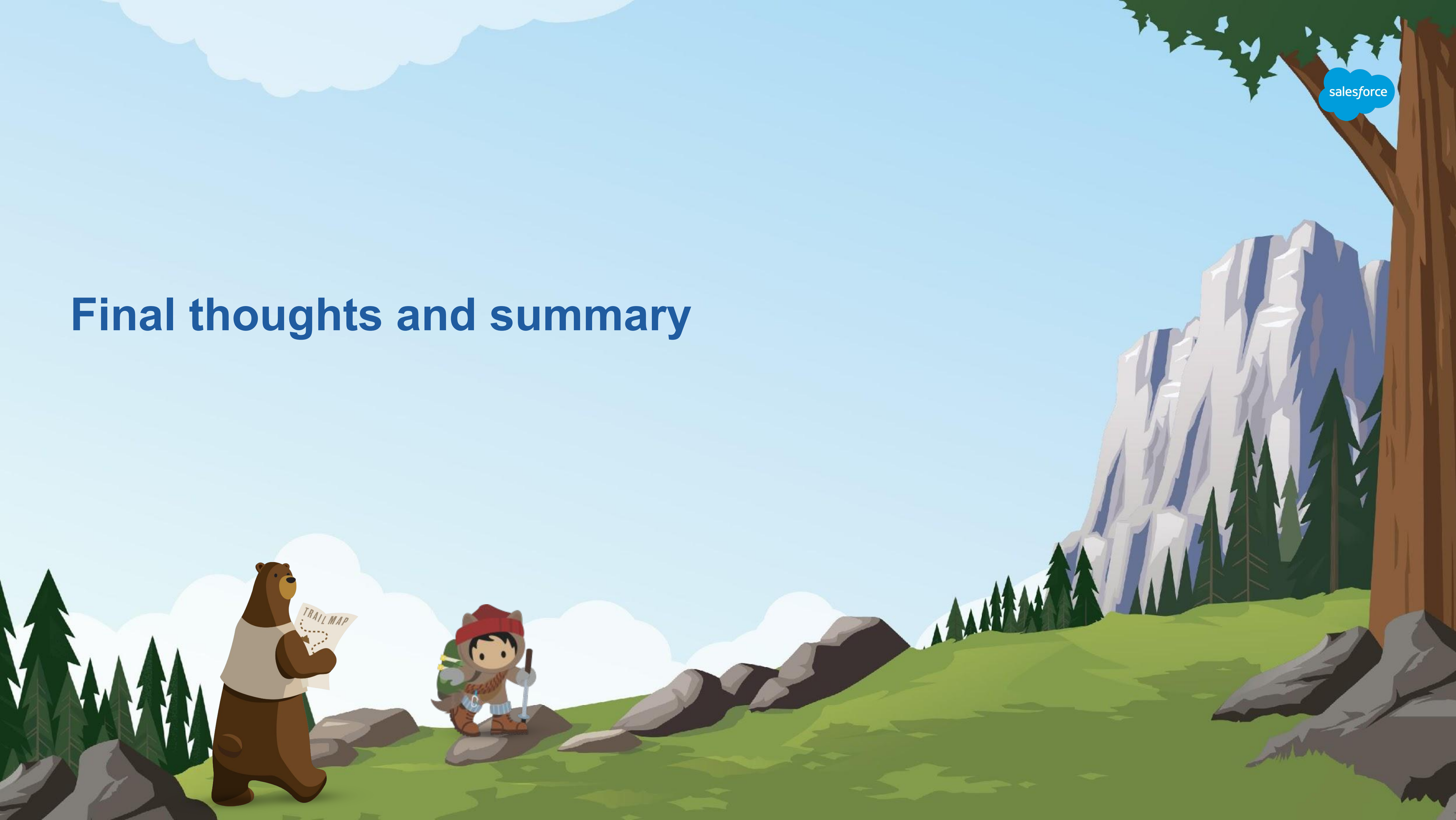
Department
mkto_si__Last_Interesting_Moment__c
Description OtherPostalCode
et4ae5__Mobile_Country_Code__c Title
mkto2__Acquisition_Program_Id__c
JigsawContactId ReportsToId OtherCity
pi__last_activity__c MailingLongitude
pi__first_activity__c AssistantPhone HomePhone
Fax OtherStreet Partner_Last_Name__c
mkto_si__Last_Interesting_Moment_Desc__c
mkto2__Acquisition_Program__c Jigsaw
Company__c OtherLongitude AssistantName
Salutation OtherLatitude Purchase_Motivation__c
Secondary_Email__c TimetoPurchase__c
mkto_si__Last_Interesting_Moment_Source__c
MailingGeocodeAccuracy MailingLatitude
pi__created_date__c CommentCapture__c
Preferred_Communication_Method__c
TopPriorityValue__c
mkto_si__Last_Interesting_Moment_Type__c
OtherState TopPriorityProcess__c OtherCountry
MasterRecordId OtherGeocodeAccuracy
TopPriorityProduct__c

emailbounceddate
lastcurequestdate lastcuupdatedate
lastreferenceddate lastvieweddate
mkto2__acquisition_date__c
mkto_si__hidedate__c pi__grade__c
pi__notes__c pi__utm_content__c
account_link_easy_closets__c
csat_survey_completed_date__c
csat_survey_net_promoter_score__c
csat_survey_results_link__c birthdate
mkto_si__last_interesting_moment_date__c
pi__campaign__c pi__comments__c
pi__first_search_term__c
pi__first_search_type__c
pi__first_touch_url__c pi__score__c
pi__url__c pi__utm_campaign__c
pi__utm_medium__c pi__utm_source__c
historical_lead_score__c pi__utm_term__c
first_activity_timestamp__c
predicted_likelihood_to_purchase_2__c

Leakers removed by
data scientist hand tuning: 42

best_time_to_call_date__
c total_lead_score__c
csat_customer_service_s
urvey_disallowed__c
referral_credit_applied__c
referral_days_til_purchas
e__c
predicted_likelihood_to_p
urchase__c createdbyid
createddate
lastactivitydate
lastmodifieddate
last_activity_date__c
systemmodstamp

Final thoughts and summary



Solve for all customers, not just one

Thresholds are tricky to choose

- What is a good feature and what is a bad leaker?

Easy to optimize for one model, but not for thousands

- Choosing a threshold that perfects one model, but makes hundreds worse is not good!

“Smart” decisions based on data shape preferred

- for example, auto-bucketizing - let the algorithm figure out a smart way

Lots of experimentation

- to learn heuristics that can be translated into algorithms

Key Takeaways

Enterprise data is very messy

- Often leads to hindsight bias/label leakage
- “Too good to be true” is a real problem

Standard Machine Learning pipeline is not sufficient

- Time based evaluation is needed to know how your models are doing
- You cannot simply optimize for best model at training time

Novel approaches needed to detect and remove leakage

- both on raw and transformed data
- choosing the right threshold to satisfy all customers

TransmogrifAI

All the methods discussed here are part of our open-source library, *TransmogrifAI*

- Built on top of SparkML
- https://github.com/salesforce/TransmogrifAI

We are hiring more data scientists!

thank you