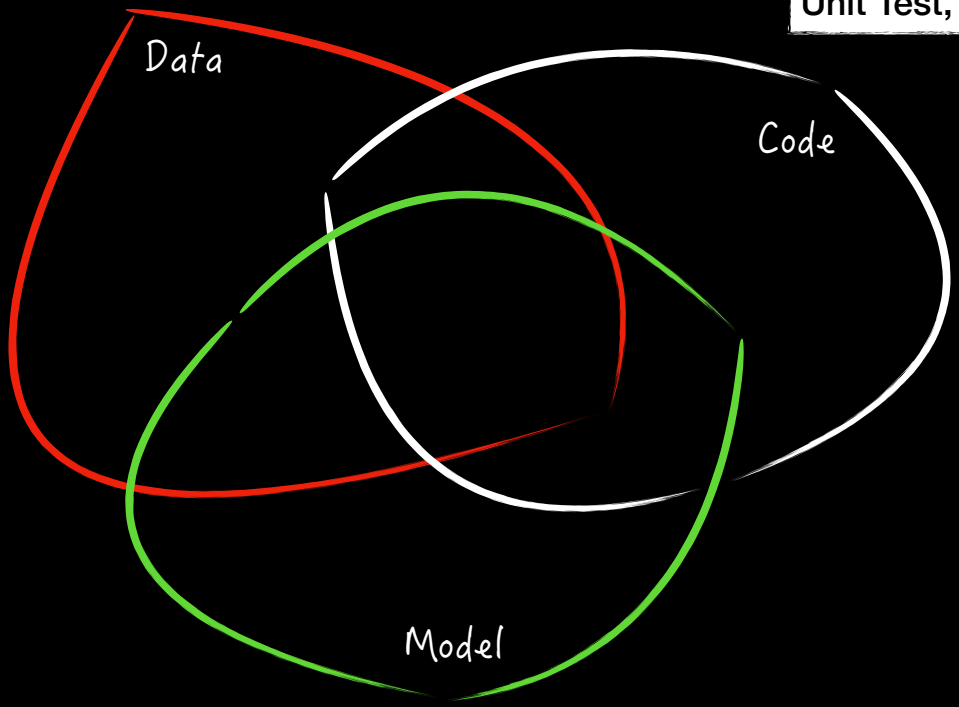
Quality Assessment

Jin Guo Oct 22th, 2020

Unit Test, Code Review for ML Code



Tests During Model Development

Test against a simpler model as a baseline.

Test the impact of each tunable hyperparameter.

Test the effect on the model's output after perturbation the input.

Perturbation Test

Test the effect on the model's output after perturbation the input

- Minimum Functionality test (MFT)
- Invariance test (INV)
- Directional Expectation test (DIR)

Capability	Min Func Test	INVariance	DIRectional
Vocabulary	Fail. rate=15.0%	16.2%	C 34.6%
NER	0.0%	B 20.8%	N/A
Negation	A 76.4%	N/A	N/A
...			

Test case	Expected	Predicted	Pass?
A Testing **Negation** with ***MFT***	Labels: negative, positive, neutral		
`Template: I {NEGATION} {POS_VERB} the {THING}.`			
I can't say I recommend the food.	neg	pos	x
I didn't love the flight.	neg	neutral	x
...			
Failure rate = 76.4%			
B Testing **NER** with ***INV***	Same pred. (inv) after removals / additions		
@AmericanAir thank you we got on a different flight to [Chicago → Dallas].	inv	pos / neutral	x
@VirginAmerica I can't lose my luggage, moving to [Brazil → Turkey] soon, ugh.	inv	neutral / neg	x
...			
Failure rate = 20.8%			
C Testing **Vocabulary** with ***DIR***	Sentiment monotonic decreasing (↓)		
@AmericanAir service wasn't great. You are lame.	↓	neg / neutral	x
@JetBlue why won't YOU help them?! Ugh. I dread you.	↓	neg / neutral	x
...			
Failure rate = 34.6%			

Figure 1: CheckListing a commercial sentiment analysis model (G). Tests are structured as a conceptual matrix with capabilities as rows and test types as columns (examples of each type in A, B and C).

Ribeiro, Marco Tulio, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList."

Tests During Model Development

Test against a simpler model as a baseline.

Test the impact of each tunable hyperparameter.

Test the effect on the model's output after perturbation the input.

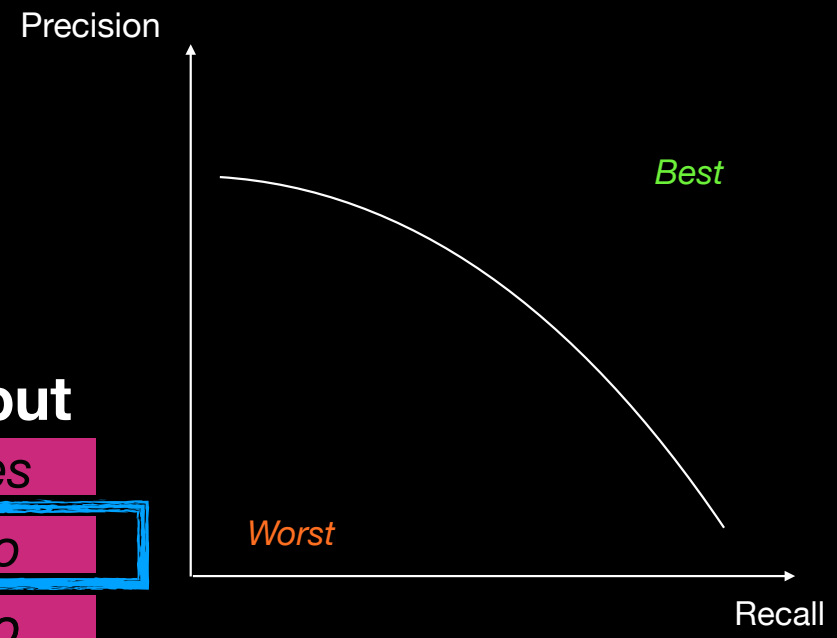
Test the model for implicit bias.

Test model quality on important data slices.

Data slicing

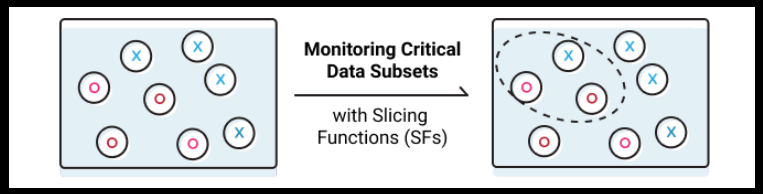
A subset that is particular relevant for the project/feature objectives.

Input	Actual Output	Model Output
1	No	0.8 -> Yes
2	No	0.7 -> No
3	Yes	0.75 -> No
.....	.....	.....
14	Yes	0.4



Data slicing

A subset that is particular relevant for the project/feature objectives.



snorkel

```
from snorkel.slicing import slicing_function


@slicing_function()
def short_link(x):
    """Return whether text matches common pattern for shortened ".ly" links."""
    return int(bool(re.search(r"\w+\.ly", x.text)))


scorer.score_slices(
    S=S_test, golds=Y_test, preds=preds_test, probs=probs_test, as_dataframe=True
)
```

	F1
OVERALL	0.925000
SHORT_COMMENT	0.666667
KEYWORD_PLEASE	1.000000
REGEX_CHECK_OUT	1.000000
SHORT_LINK	0.500000
TEXTBLOB_POLARITY	0.727273

https://www.snorkel.org/use-cases/03-spam-data-slicing-tutorial

Vincent S. Chen, Sen Wu, Zhenzhen Weng, Alexander Ratner, Christopher Ré
“Slice-based Learning: A Programming Model for Residual Learning in Critical Data Slices”

Tests During Model Development

Test against a simpler model as a baseline.

Test the impact of each tunable hyperparameter.

Test the effect on the model's output after perturbation the input.

Test the model for consideration of inclusion.

Test model quality on important data slices.

Test the data pipeline has appropriate privacy controls.

Test the relationship between offline proxy metrics and the actual online impact metrics.

Tests During Model Development

Test against a simpler model as a baseline.

Test the impact of each tunable hyperparameter.

Test the effect on the model's output after perturbation the input.

Test the model for implicit bias.

Test model quality on important data slices.

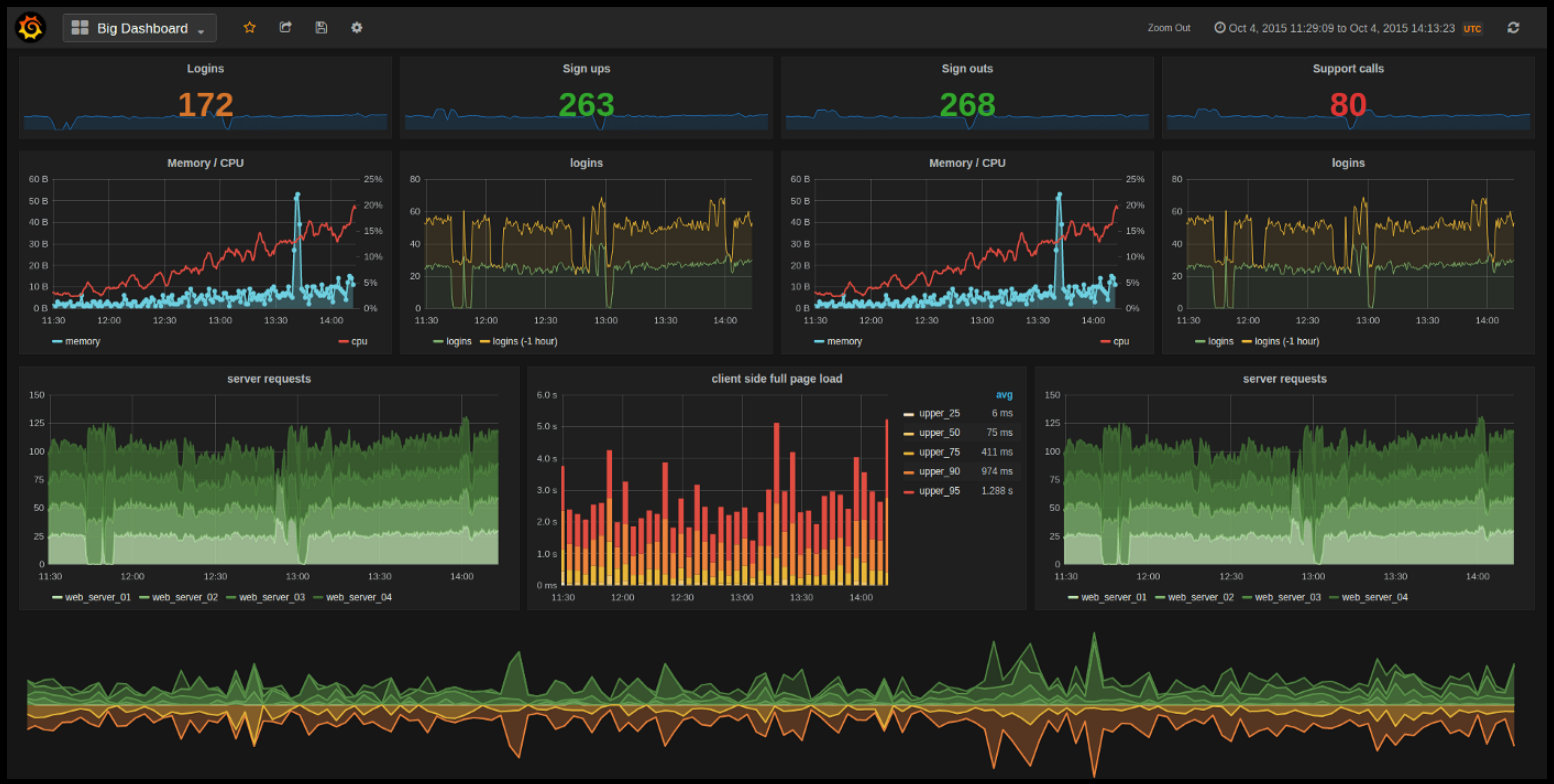
Test the relationship between offline proxy metrics and the actual online impact metrics.

Telemetry Design

Responsible for collecting observations about how users are interacting with the system

- Monitoring system works correctly
- Understand the impact on users
- Gather new training data

Monitoring



https://medium.com/@pacroy/application-telemetry-with-prometheus-sap-blogs-c4b5b6239d28

Understand User Impact

To determine if users are getting positive or negative outcomes and if the system is achieving its goals.

- Which experiences do users receive and how often do they receive them?
- What actions do users take in each experience?
- What experiences tend to drive users to look for help or to undo or revert their actions?
- What is the average time between users encountering a specific experience and leaving the application?
- Do users who interact more with the intelligent part of the system tend to be more or less engaged (or profitable) over time?

Activity

- Group 1: Amazon: Shopping app feature that detects the shoe brand from photos;
- Group 2: Google: Tagging uploaded photos with friends' names;
- Group 3: Spotify: Recommended personalized playlists;
- Group 4: Microsoft: Code completion recommendation in IDE.

- What information should the telemetry system capture?
- How are you going to use the information to identify and debug potential problems?
- How costly is it to collect the data? How do you plan to manage the cost?
- Any challenges/risks for the your telemetry design?