Model Evaluation For Humans

About me

Husband and Father
Philly data scientist
Wrote a haiku once

Disclaimers

- Most useful for new(er) data scientists
- But there may be take aways for everyone
- Oversimplication
- Guidelines/examples are not exhaustive
- Breadth not depth
- Going to go fast
- Sometimes the obvious needs to be stated

Goal of this talk

Discuss guidelines and practices for robust model evaluation

Outline

- Motivation
- Before development
    - Business & Data Understanding
- During development
    - Data Preparation → Modeling → Evaluation
- Deployment and beyond
    - Deployment → Evaluation & Monitoring → Iteration

Motivation: all models are wrong

"All models are wrong, ..."

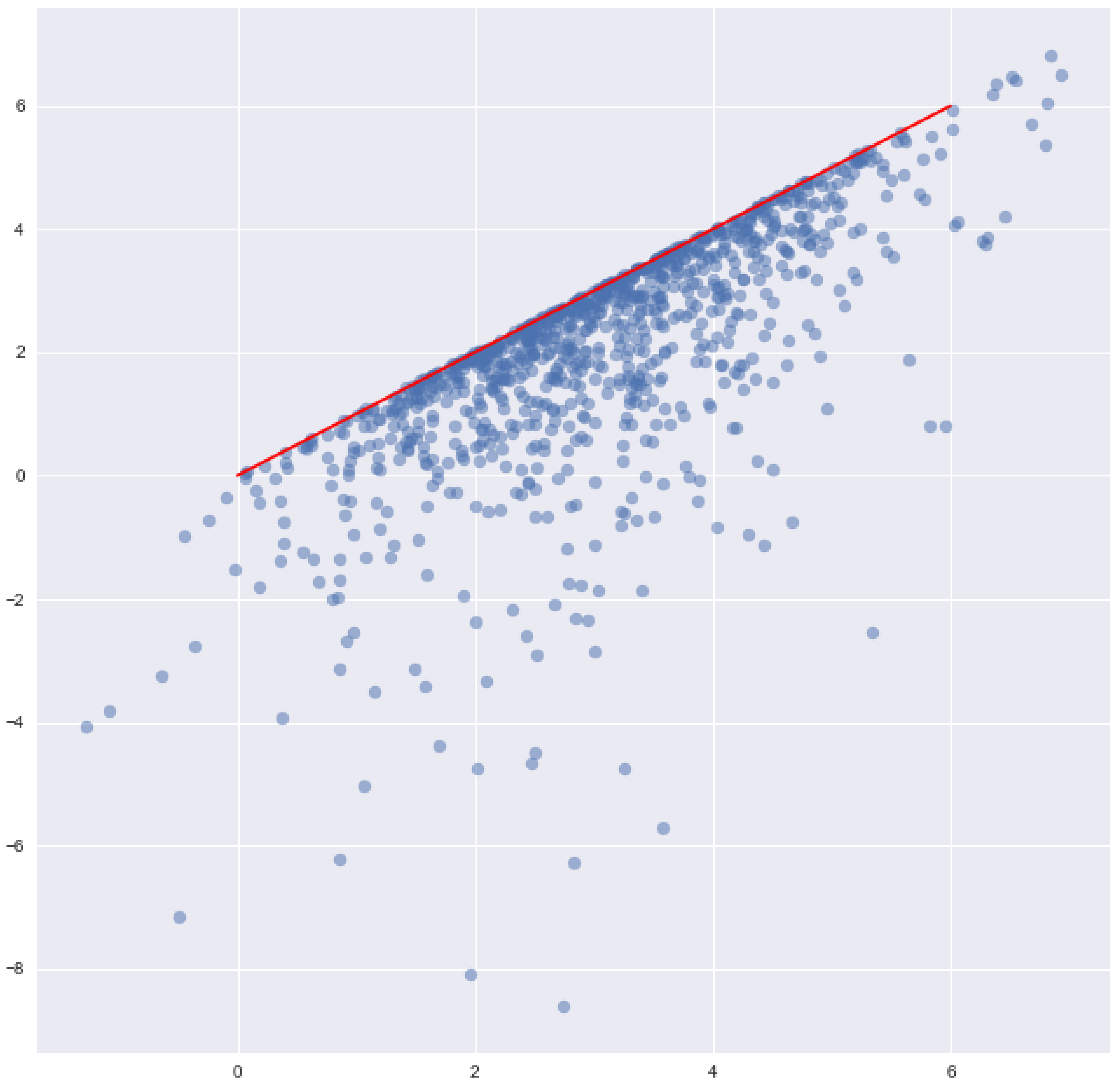
$$\underset{\theta}{\operatorname{argmin}}\ L(y, \hat{y})$$

$$MLE := \underset{\theta}{\operatorname{argmax}}\ \pi(D \mid \theta)$$

$$\hat{y} = f(x) + \epsilon$$

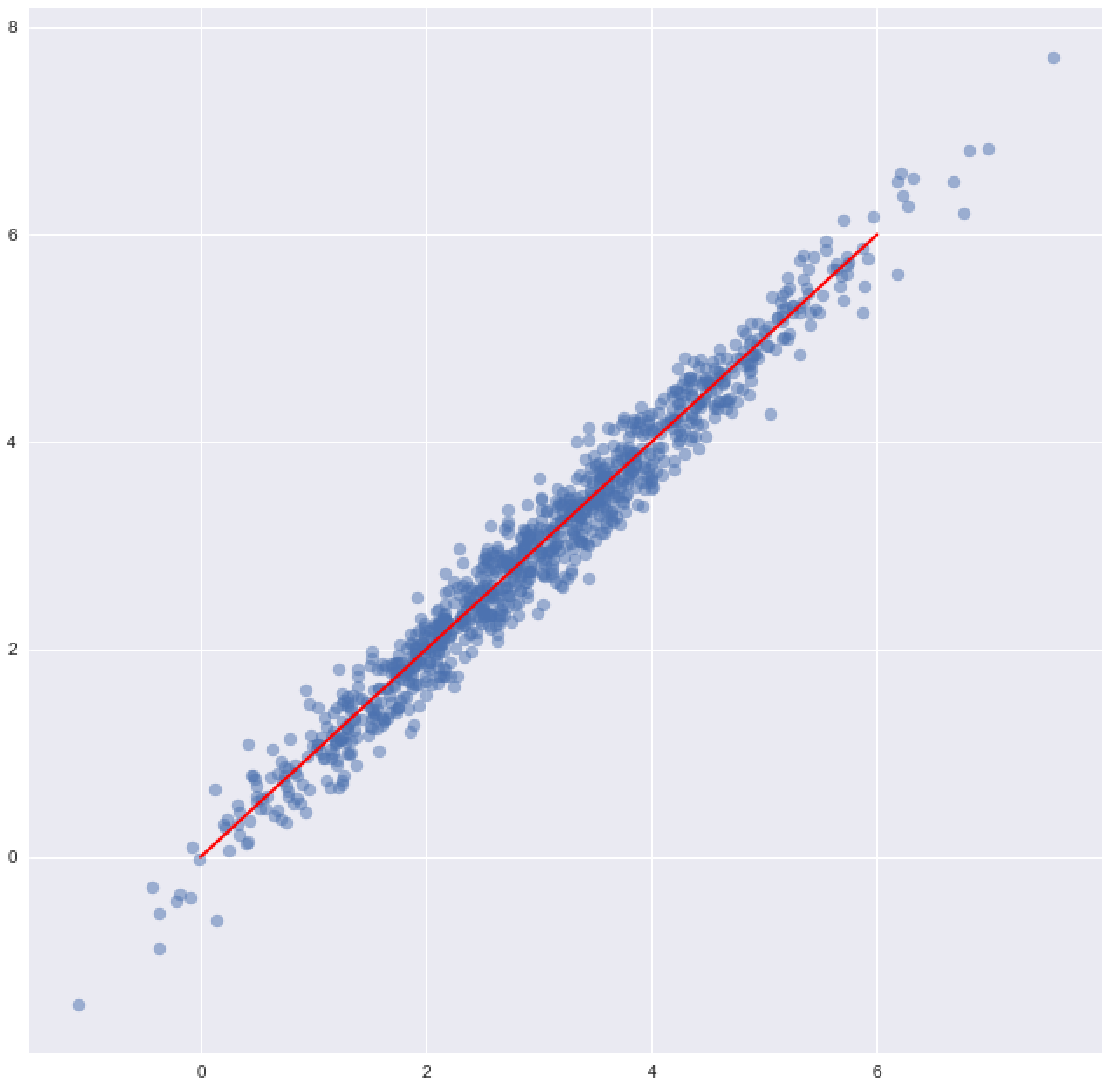
$$P(\theta \mid D) \propto P(D \mid \theta)P(\theta)$$

Ask me how I know…



6
4
2
0
-2
-4
-6
-8
0
2
4
6

On a better day… still wrong


8
6
4
2
0
0
2
4
6

The zen of model evaluation

"*All models are wrong, some are useful*"

The zen of model evaluation summarized

- All models are wrong
- Some are useful
- Others are not
- Usefulness counts

Before development

- Highly context specific
- Essential for rest of project

Hazards of getting this wrong

- Model does not solve problem of interest
- Model indirectly solves problem of interest
- Model cannot be evaluated in prod
    - Reduced confidence in model
- Diminishing returns
- Iteration is impaired

Some things to think about…

Think about framing the problem

- How will this model be evaluated?
- How will success be *measured*?
- Is this a classification or regression problem?
- Choose metrics and KPIs carefully
    - Conversion to business value or terminology
    - Comparability
        - Across time, data, iterations, etc.

Think about deployment and future iterations

- "Nothing is more permanent than temporary"
  - Prepare for deployment
  - Prepare for future iterations
- Example: A/B testing

Think about the feedback loop

- How will outcomes from the model be collected?
- Example: Ad recommendations
- Example: Intervention for hospital readmission
    - Is the model that bad or that good?

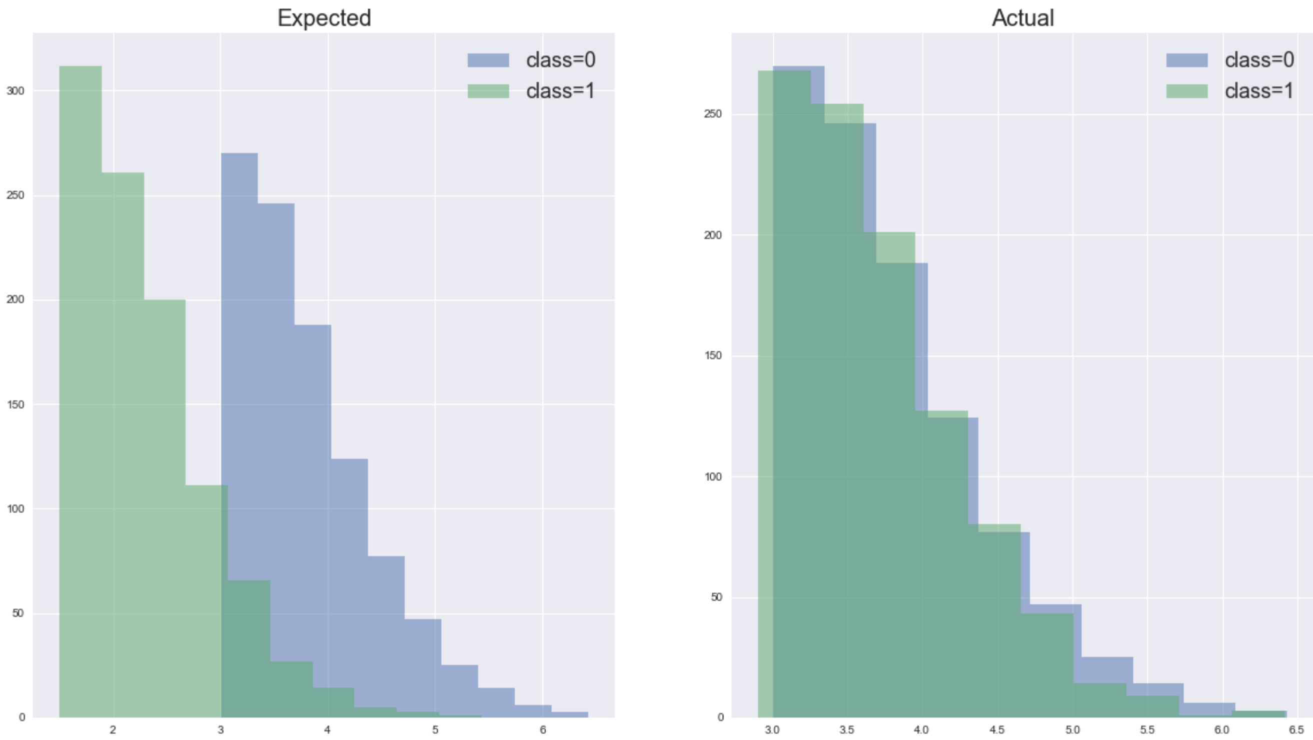
During development

- Most generalizable
- Opportunity adopt and standardize best practices

Test assumptions

- $f(\quad) =$
- Check distributions of data: Plot, plot, plot
- Check features against conventional wisdom
- Etc.

Testing assumptions: Example



Establish a baseline

A good baseline is

- Easy to implement
- Easy to understand
- Simple but not a strawman

Examples of baselines

- A simple heuristic
  - Example: Persistency
  - Example: Ratings

- An interpretable model
  - Example: Comorbidities

- The existing model or methodology

- Acceptance criteria

Cross validation

Replicate the deployment pattern

- Avoid leakage from future examples
- More like backtesting, less like K-fold CV
  - Train on Jan. predict on Feb.
  - Train on Feb. predict on Mar.

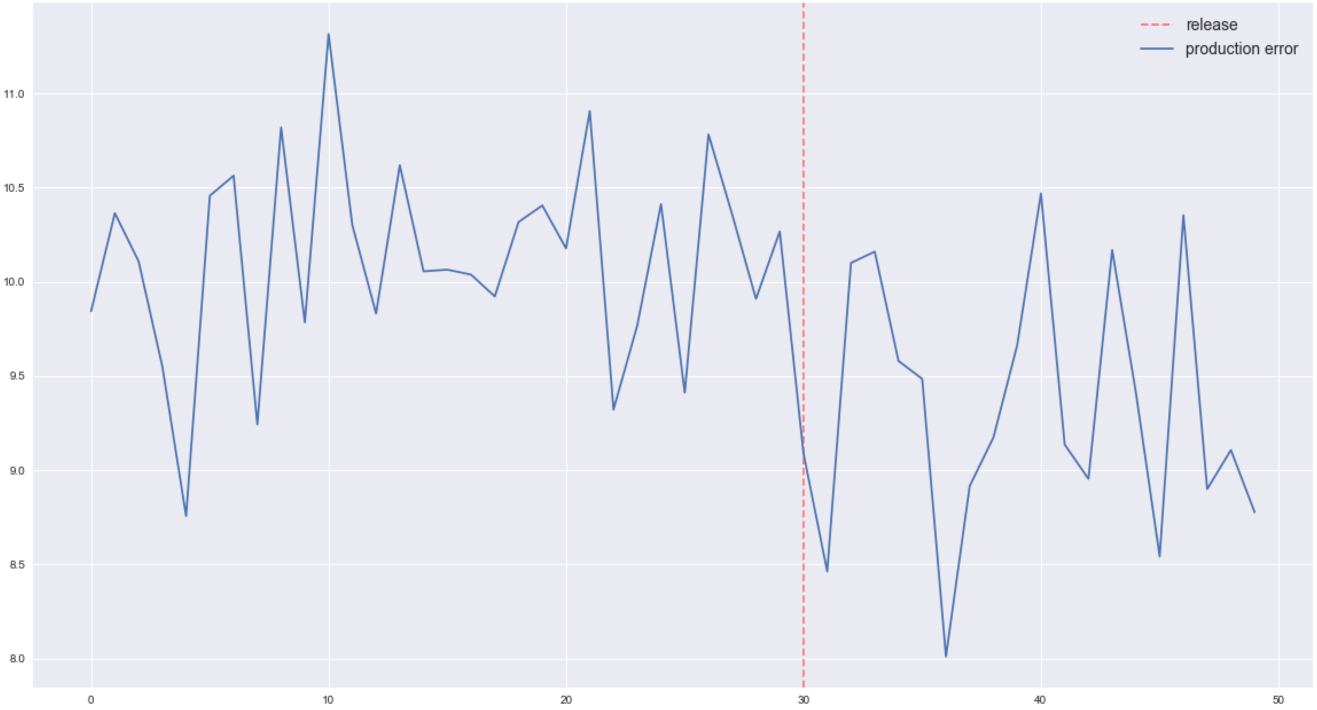
Train on a cleaner data set

- Validate modeling approach apart from noise in data set
    - Example: Ratings

Deployment and beyond

- This is what really matters (Remember the Zen!)
- Ensure that your model is (and stays) useful!
- Iterate with confidence!

Why A/B testing?



Monitoring:

- Keep it simple and actionable
  - Example: Favor Range over KL Divergence (http://stevenwhang.com/tfx_paper.pdf)

- Reuse sanity checks from development

Conclusion

- Hopefully you have some take aways
- Share your suggestions!
- https://dantegates.github.io/2019/01/07/model-evaluation-for-humans.html (https://dantegates.github.io/2019/01/07/model-evaluation-for-humans.html)

Questions?