Validate and Monitor Your Machine Learning Models

Olivier Blais

To access some useful tools:

https://github.com/moovai/model_validation_tutorial

Objective

Discuss the importance to validate properly statistical models

Share techniques to validate and monitor complex machine learning models

Notes

This session is about industry agnostic validation methods

No official standards exists to validate complex models

Only some techniques will be discussed

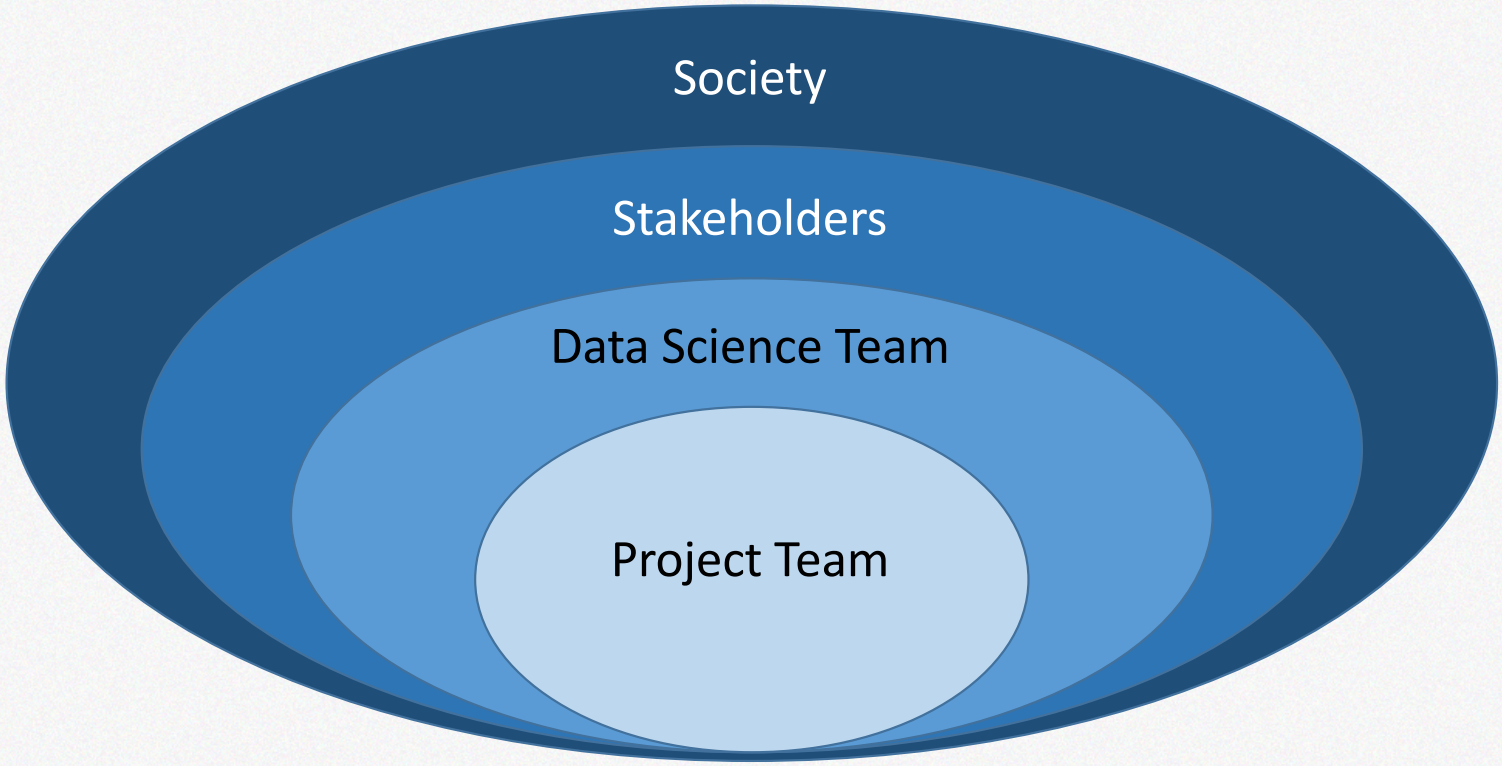
Agenda

1. Introduction
2. Validation Methodology
3. Validation tools
   a. Evaluate model performance
   b. Evaluate model stability
   c. Evaluate model biases
   d. Evaluate model sensitivity
4. Monitoring Methodology
5. Monitoring tools
   a. Evaluate data drift
   b. Evaluate model degradation
   c. Evaluate label stability

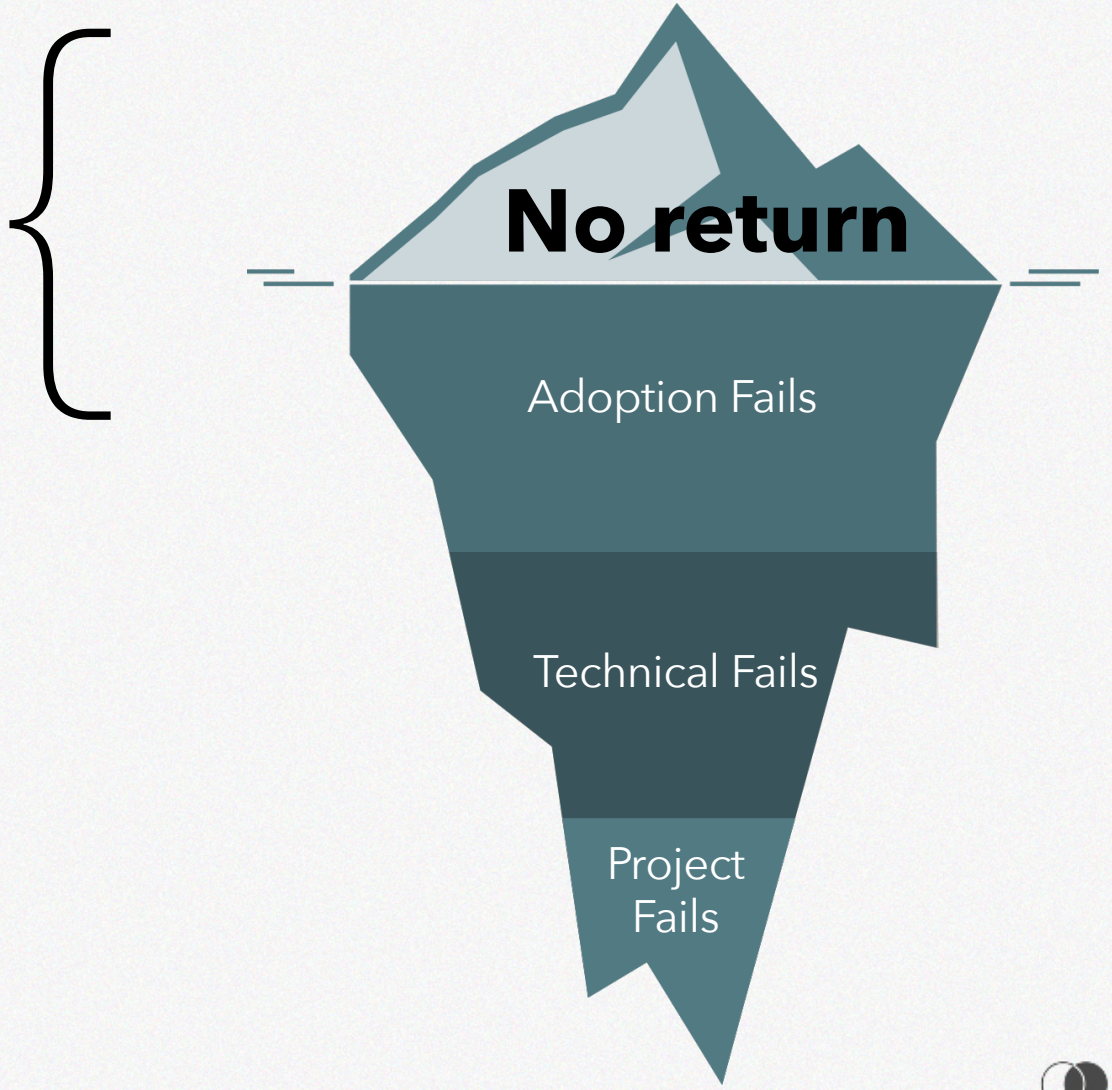
1. Introduction

87% of machine learning proof of concepts are never put in production

How Can Model Validation Help?



Iceberg of Machine Learning Failure



ODSC

MOOV AI

Examples of Project Risks

Training data quality

False positive and false negative biases

Production data accessibility

Integration to systems

User experience

Project Risks

Data

Production data is different from training data

Technology

Important class imbalance, Complex problem to optimize

Integration

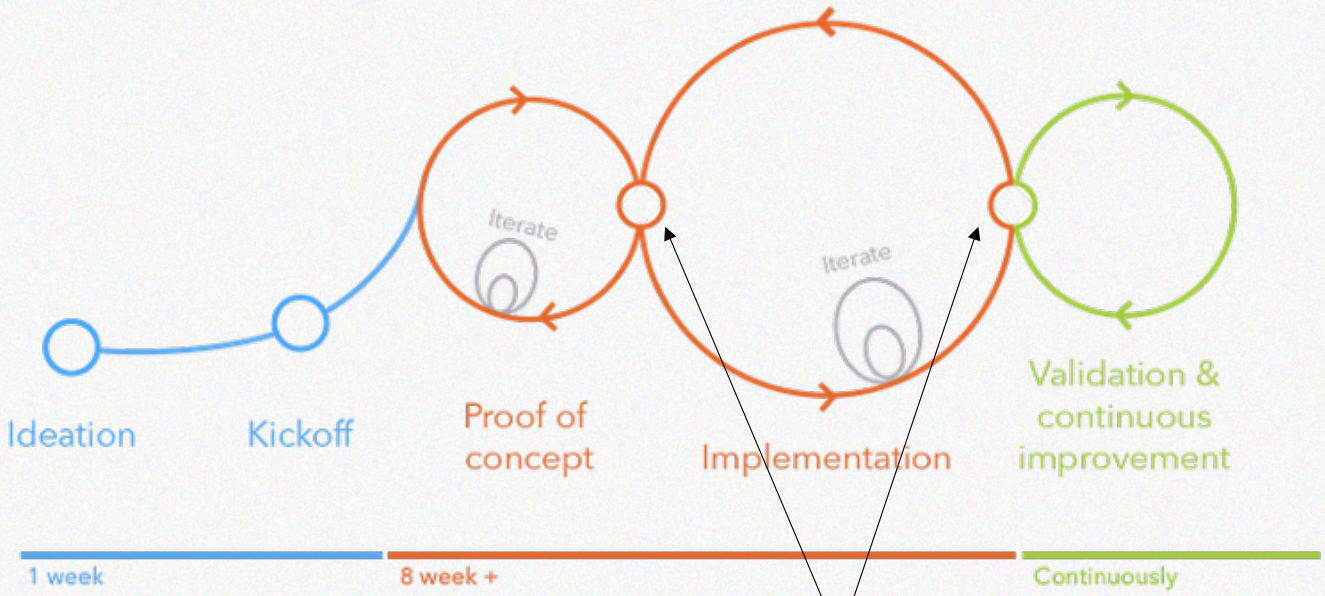
Prediction needs to be done in a different context

Adoption

Cost of false negative is important

Project Validation Process

Project Development



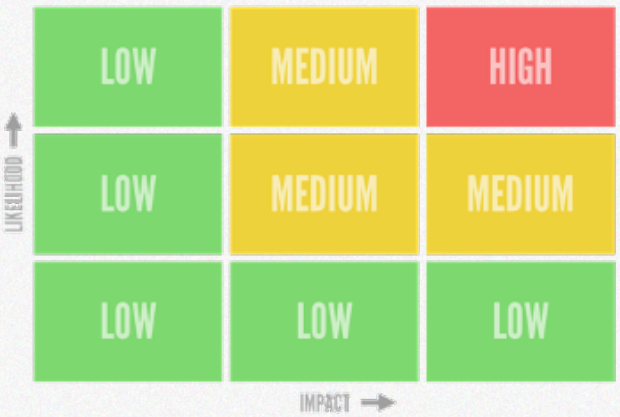
Project Validation

Define/ prevent project risks	Define/ prevent technical risks	Validation prior To deployment	On-going monitoring

ODSC

MOOV AI

Model Validation Framework

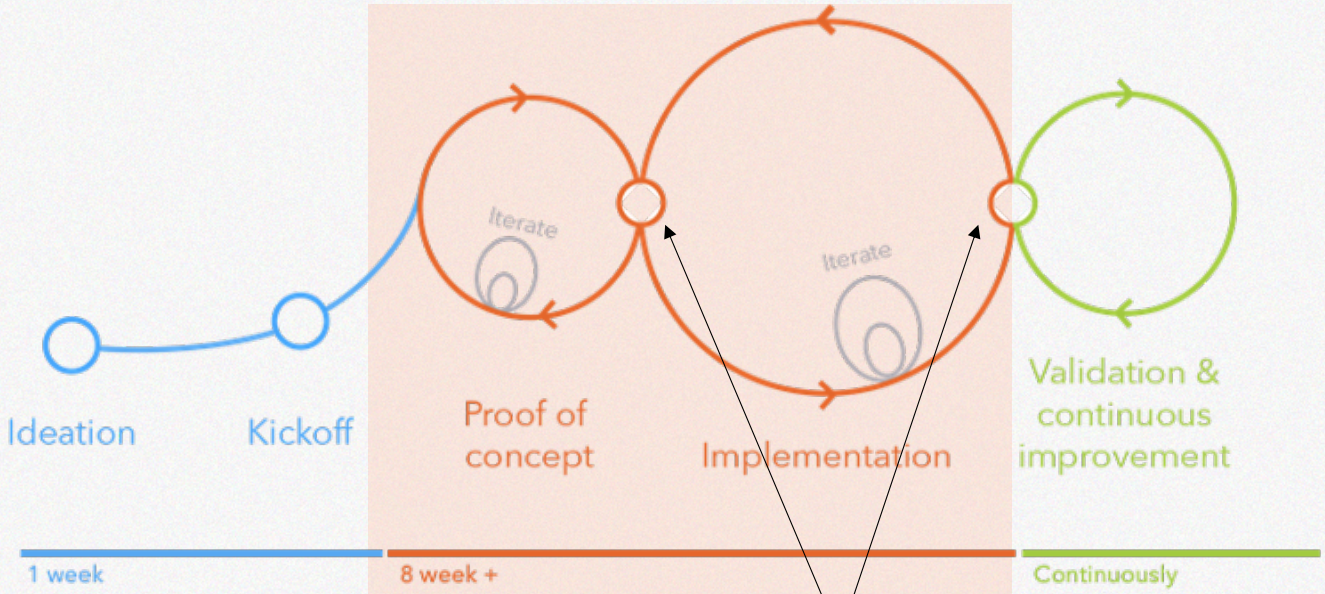


- Organizations should be obliged to validate their models
- Determines the assessment of the predictive power
- Process is carried out at least once per year (monthly)
- Both qualitative and quantitative assessment required
- Results are assessed by an independent entity

2. Validation Methodology

Initial Validation Process

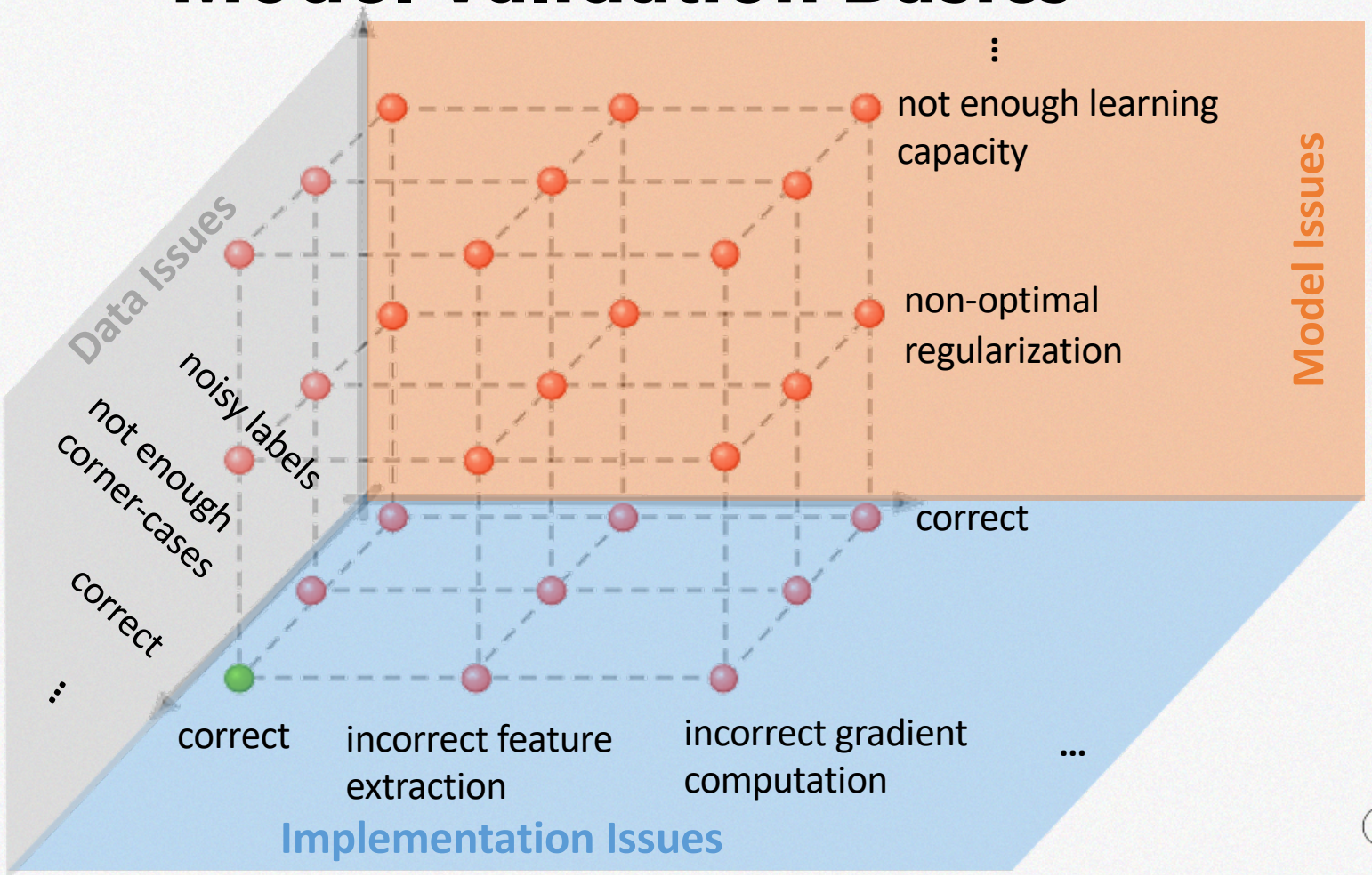
Project Development



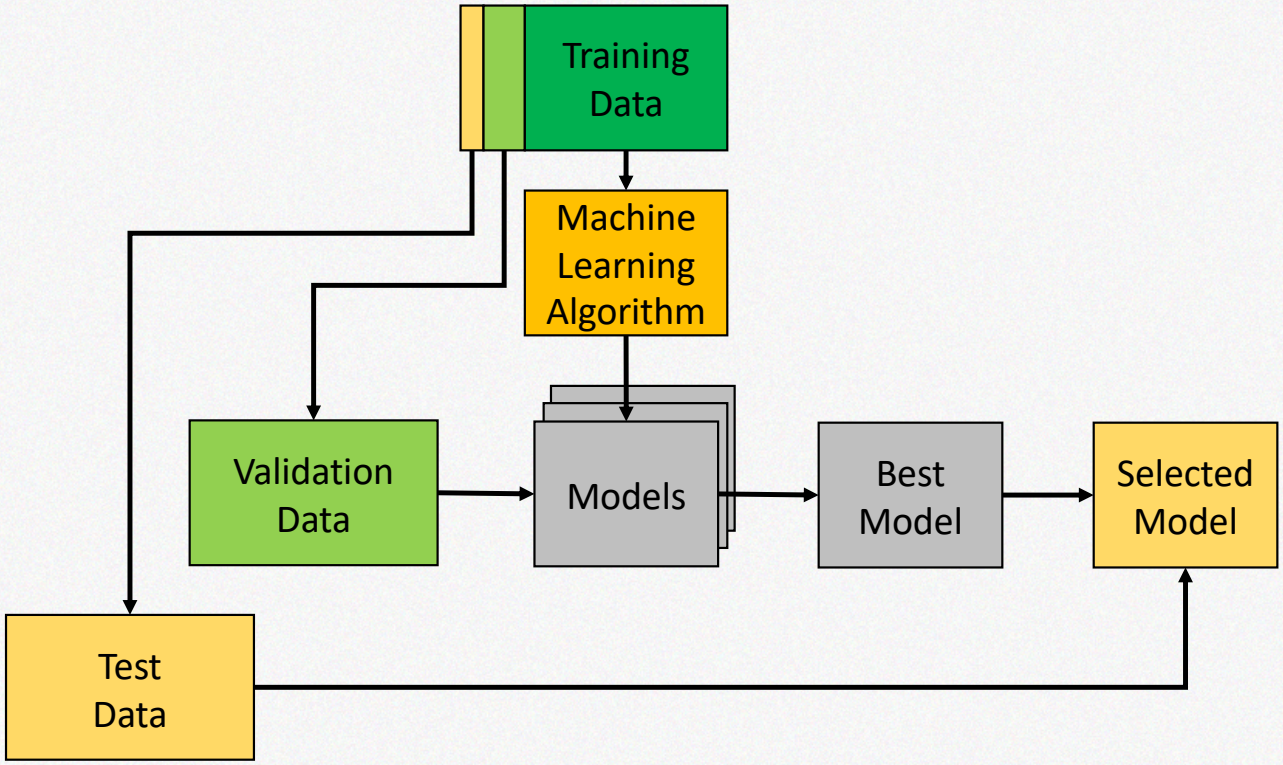
Project Validation

Define/ prevent project risks	Define/ prevent technical risks	Validation prior To deployment	On-going monitoring

Model Validation Basics



Model Validation Basics



3. Validation tools

3a) Evaluate model performance

3ai) Classification

Classification Basics

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	**Sensitivity** $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	**Specificity** $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	**Negative Predictive Value** $\frac{TN}{(TN + FN)}$	**Accuracy** $\frac{TP + TN}{(TP + TN + FP + FN)}$

Classification Basics

		Predicted Class: Positive	Predicted Class: Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error
	Negative	False Positive (FP) Type I Error	True Negative (TN)

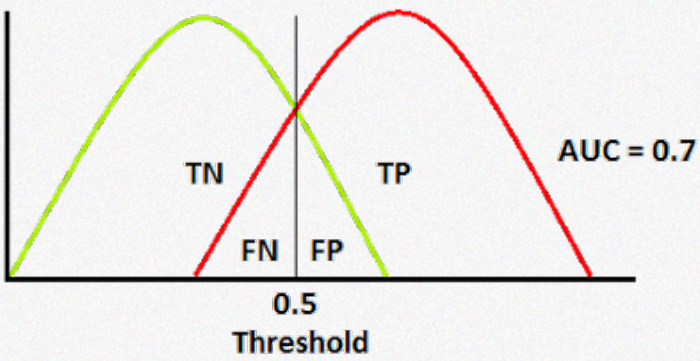
Strategies to maximize model evaluation

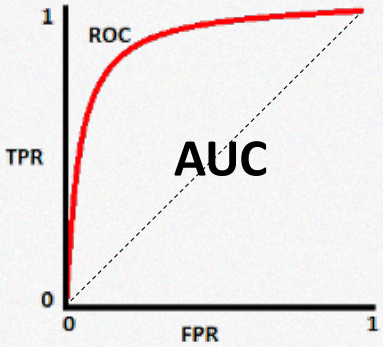
- **Measure return on investment**
  - Estimate value for each outcome
  - Measure volume of actual outcomes by class
- **Define baselines**
  - Dummy classifier
  - Previous version
  - Manual performance

Area Under the Curve (ROC-AUC)

How much model is capable of distinguishing between classes

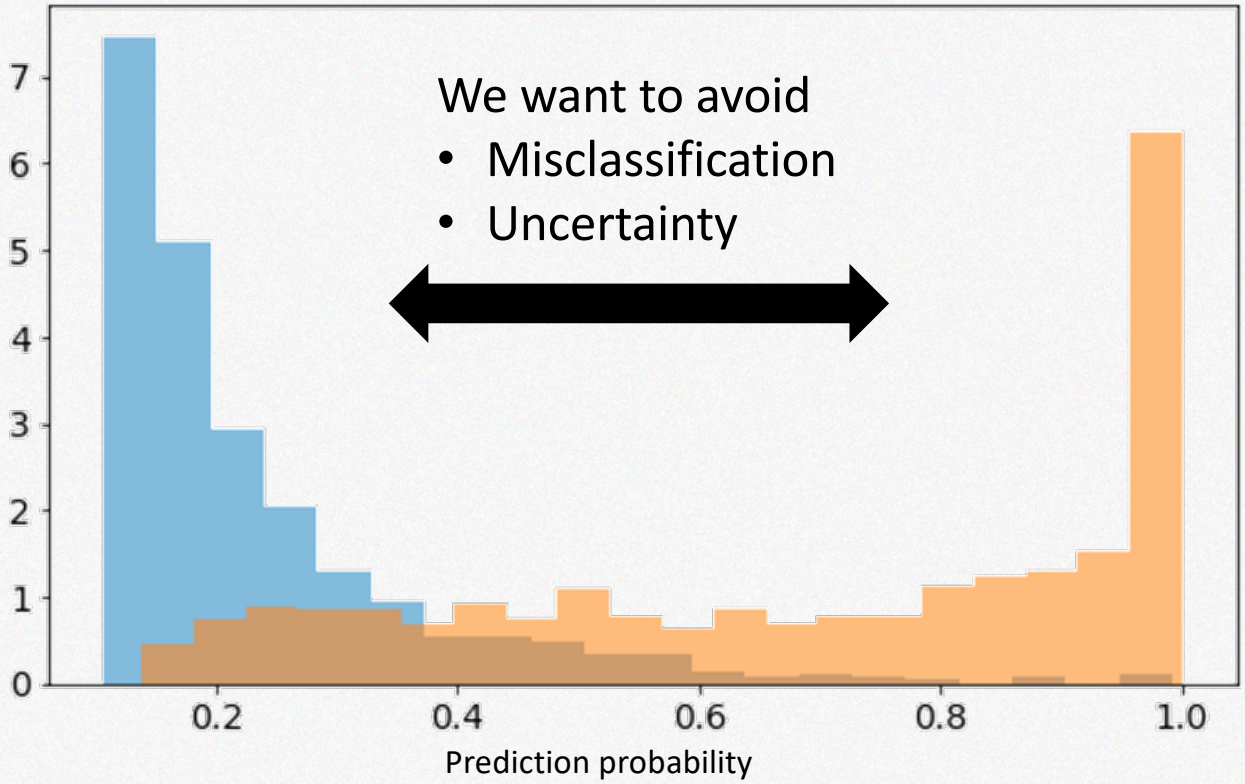
Rule of thumb
AUC = 0.5: No discrimination capacity to distinguish between positive and negative class





https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

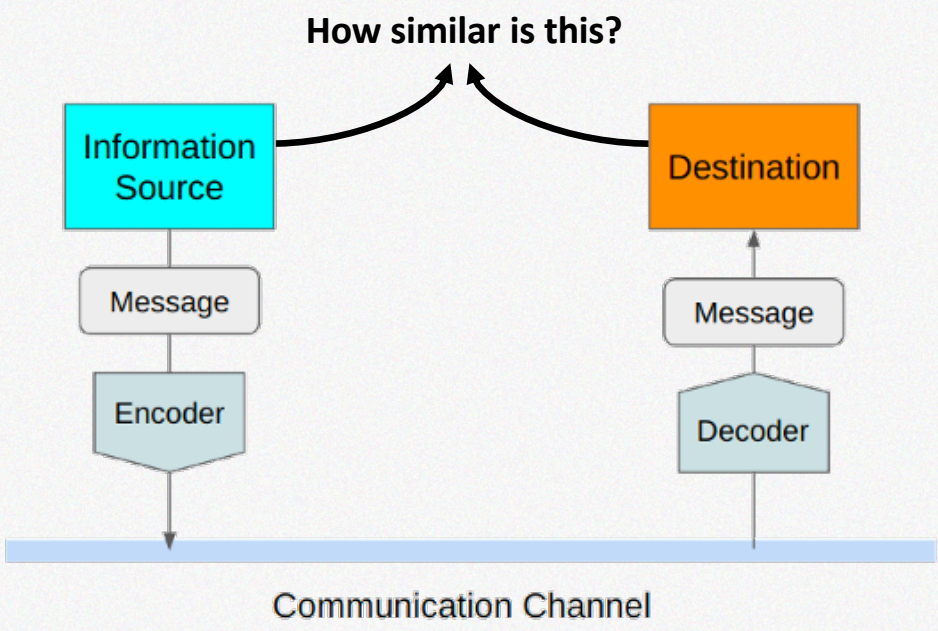
Cross-Entropy



ODSC

MOOV AI

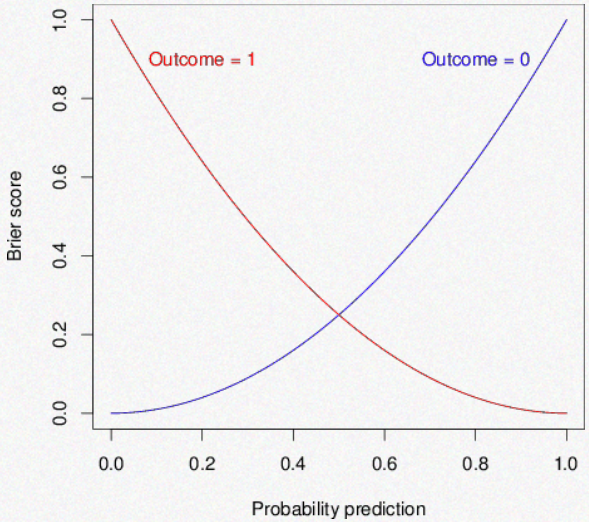
Cross-Entropy

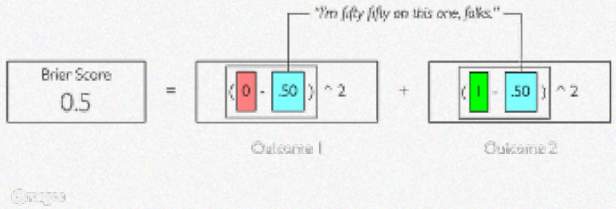


Cross Entropy (log loss) measures deviation from the truth

Rule of thumb
Cross-Entropy = 0.0: Perfect probabilities
Cross-Entropy < 0.1: Moderate uncertainty
Cross-Entropy > 0.3: Great uncertainty

Brier Score



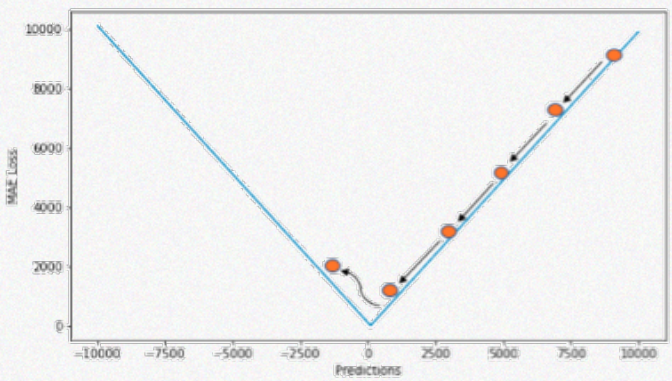


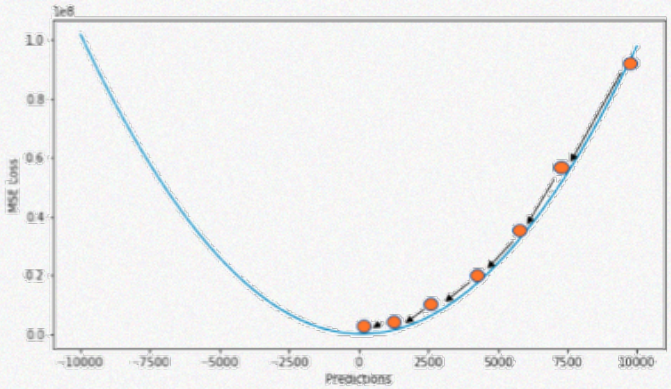
mean squared error between predicted probabilities and the expected values

Rule of thumb
Brier Score = 0.0: Perfect calibration
Brier Score = 1.0: Imperfect calibration

3aii) Regression

MSE vs MAE

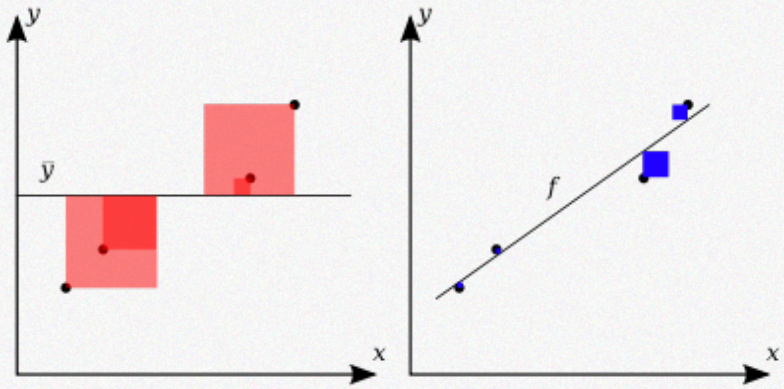




MAE (Mean absolute error): difference between the original and predicted values extracted by averaged the absolute difference over the data set.

MSE (Mean Squared Error): difference between the original and predicted values extracted by squared the average difference over the data set.

Adjusted R Squared



$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Understand the percentage of variance in the target variable explained by the model.

Rule of thumb
R-squared < 0.5: weak or low effect size
R-squared < 0.7: Moderate effect size,
R-squared > 0.7: Strong effect size

AIC or BIC?

Accuracy Term

BIC $log(P(D \mid M)) \approx \boxed{log(P(D \mid M, \hat{\theta}_M))} - \boxed{\frac{n}{2} \log m}$

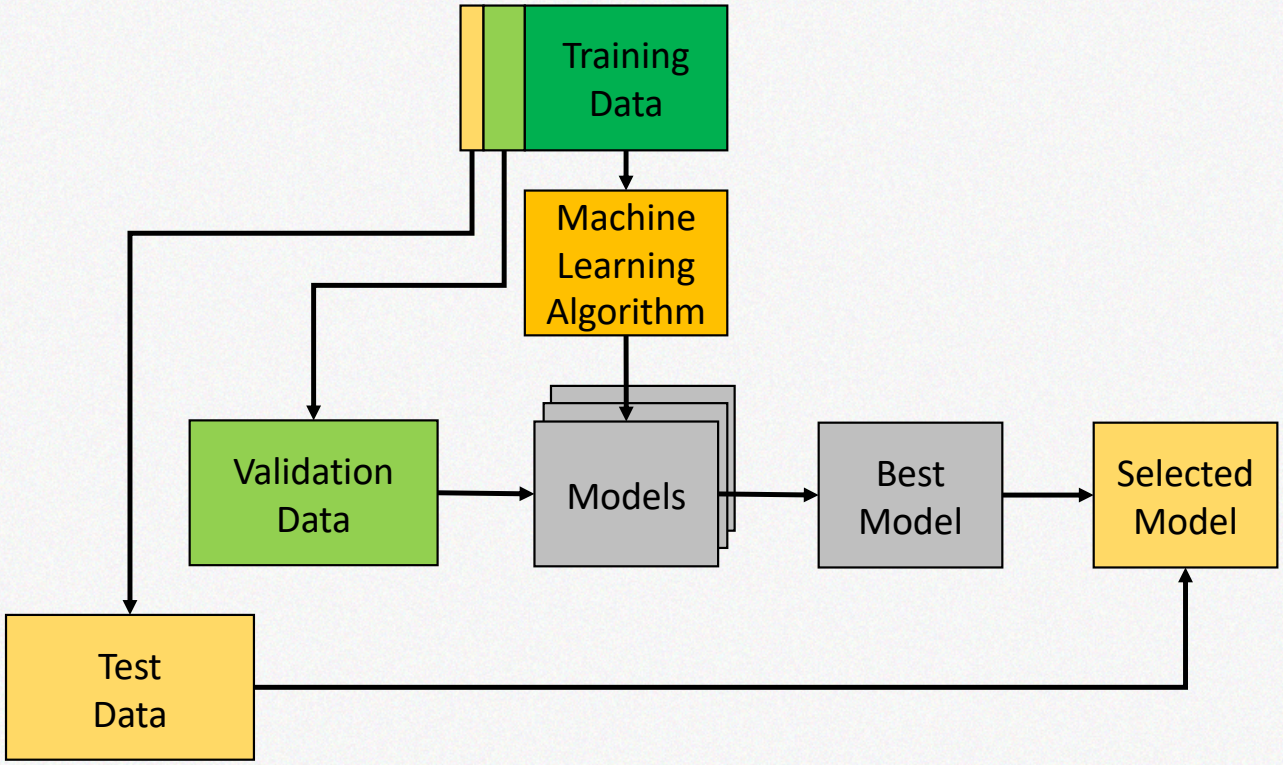
AIC $log(P(D \mid M)) \approx \boxed{log(P(D \mid M, \hat{\theta}_M))} - \boxed{n}$

Complexity Term

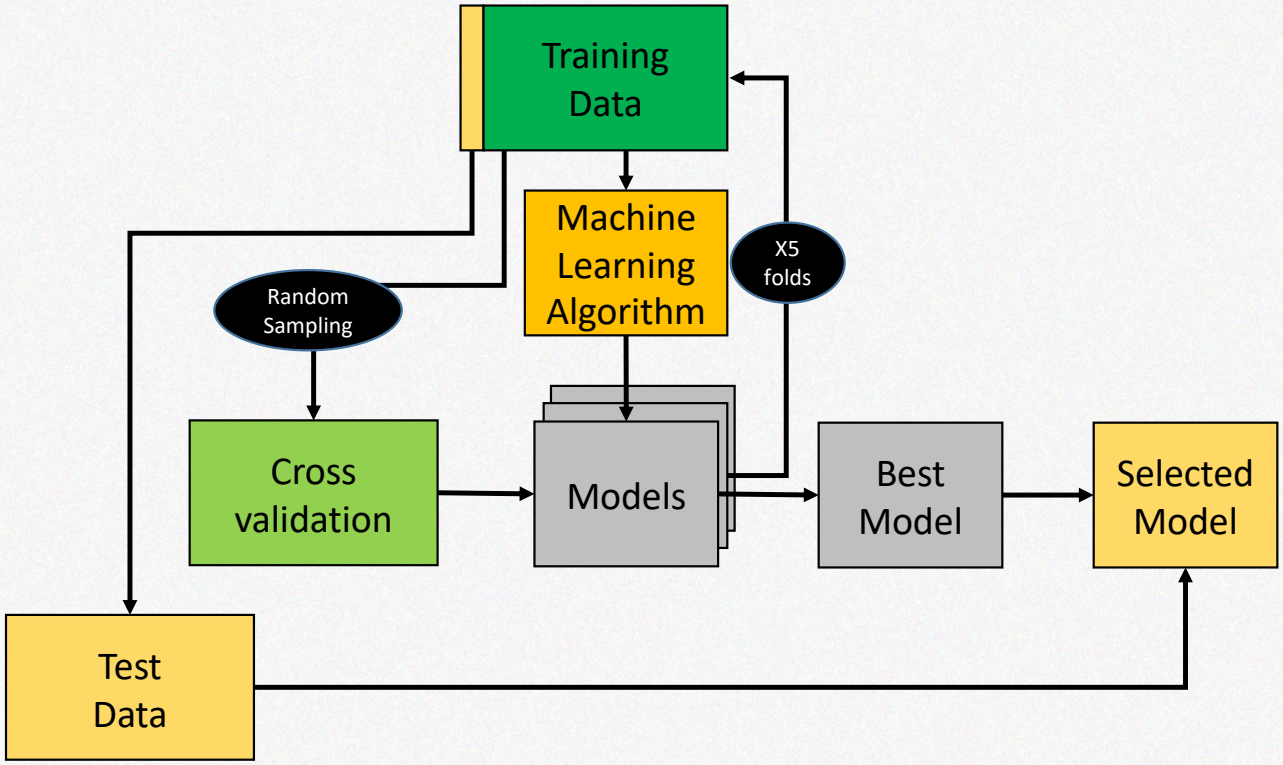
Can be still useful but cross validation is a better approach to estimate out-of-sample prediction error

3b) Evaluate model stability

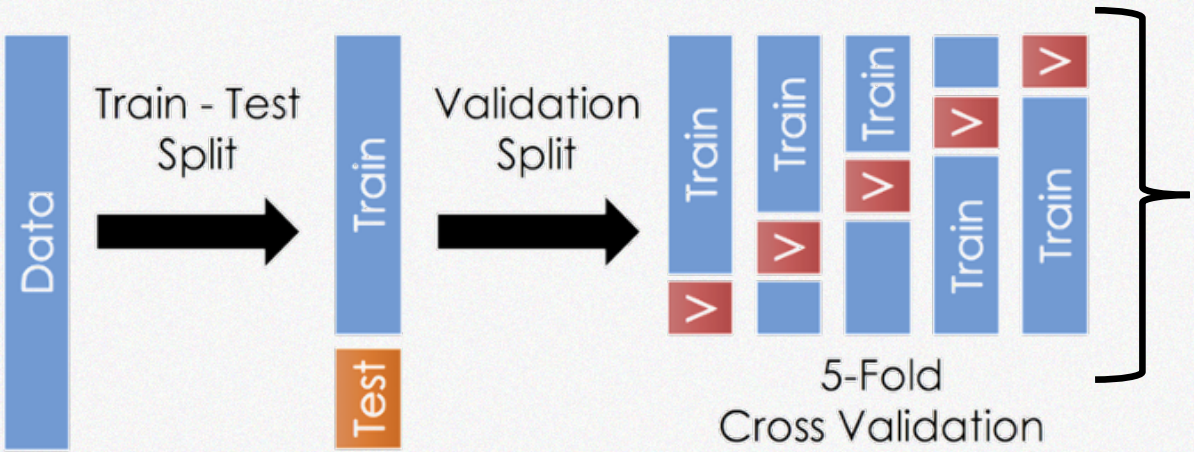
Hold-out Validation



Cross-Validation



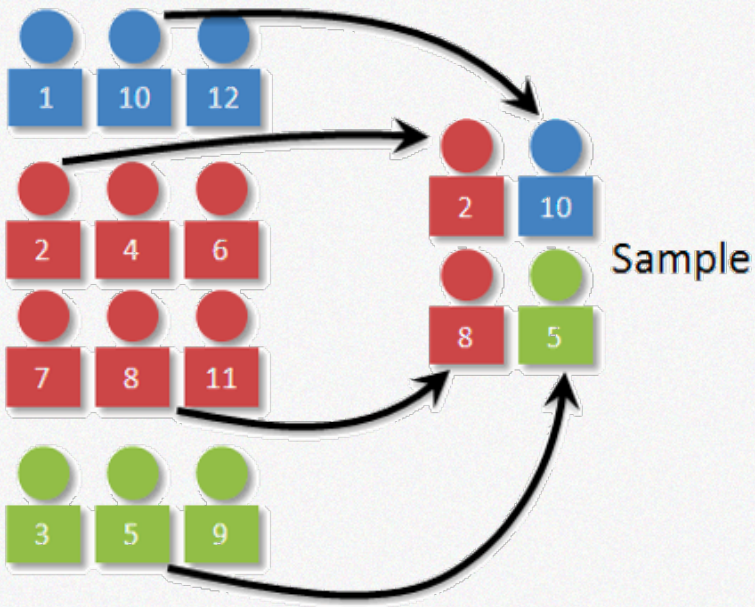
Cross-Validation



New take on existing metrics

- Mean: How good?
- Min: What's the worst case?
- Standard Deviation: How stable?

Stratified Sampling

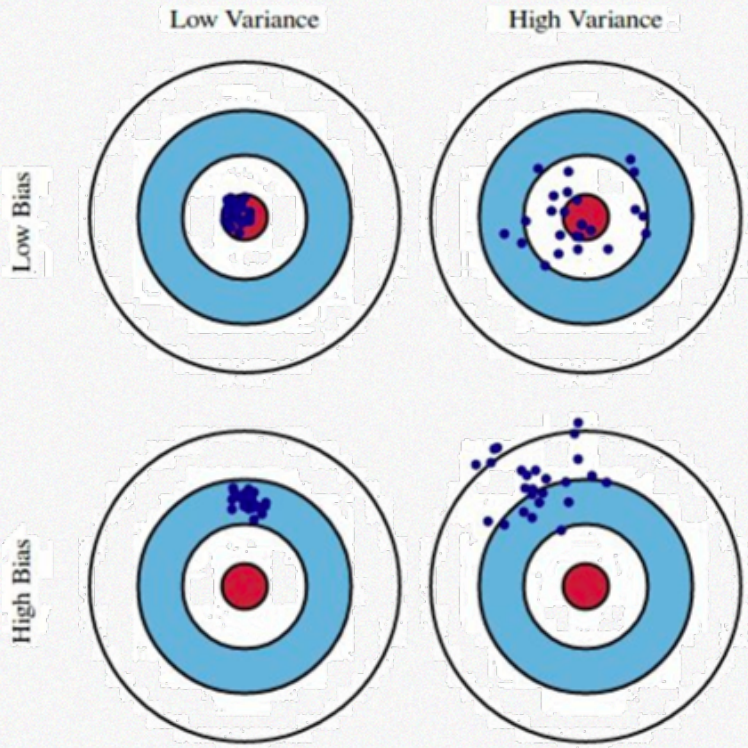


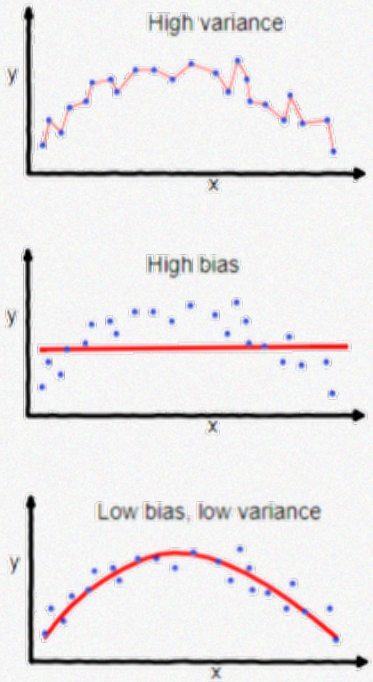
Sometimes, random sampling is not ideal

When to use stratified sampling?

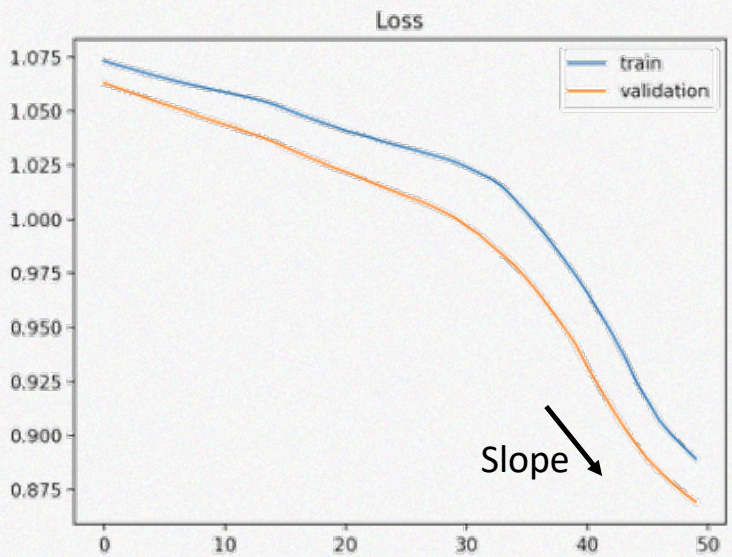
- Rare classes
- Multiple classes
- Small data
- High variance (using clustering techniques)

Bias vs. Variance

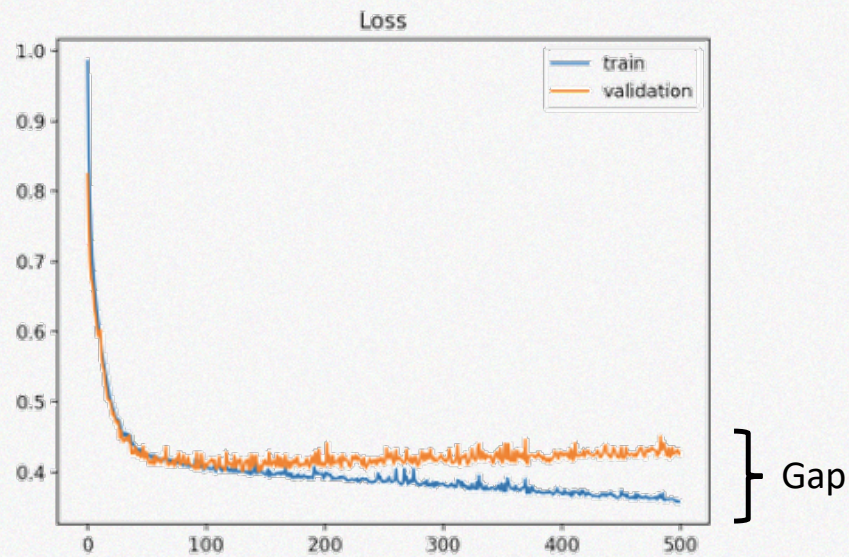




Model Fit Assessment



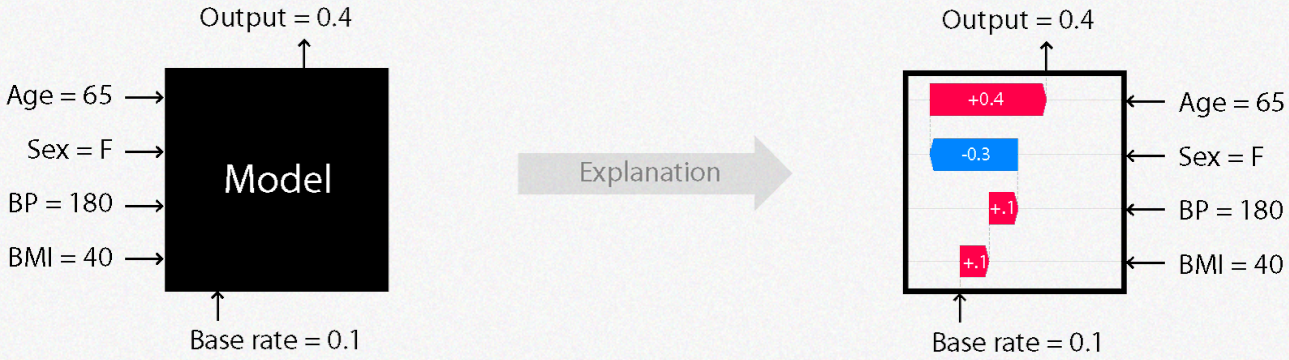
Learning curves exhibiting **underfitting** behavior



Learning curves exhibiting **overfitting** behavior

3c) Evaluate model biases

SHapley Additive exPlanations



Pros

- Interpret both local and global distributions
- Allow feature dependence on some implementations
- Shapley values can be analyzed and clustered
- Can be extracted during inference phase
- Many great graphs built-in

Cons

- Super slow to run on big datasets
- No feature dependence on KernelSHAP
- Shapley values not always simple to understand

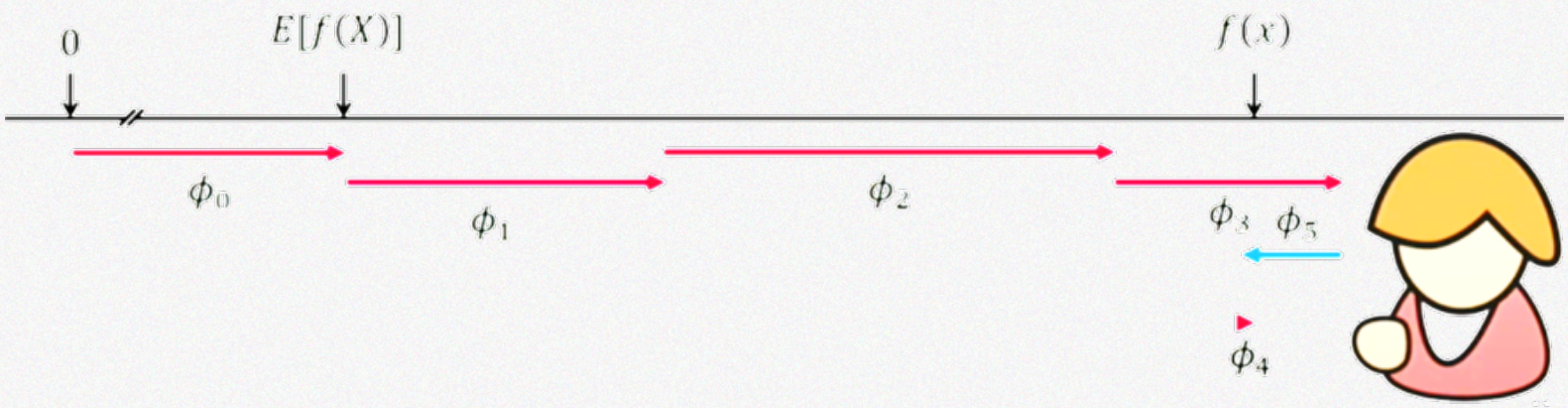
https://christophm.github.io/interpretable-ml-book/shap.html

SHapley Additive exPlanations

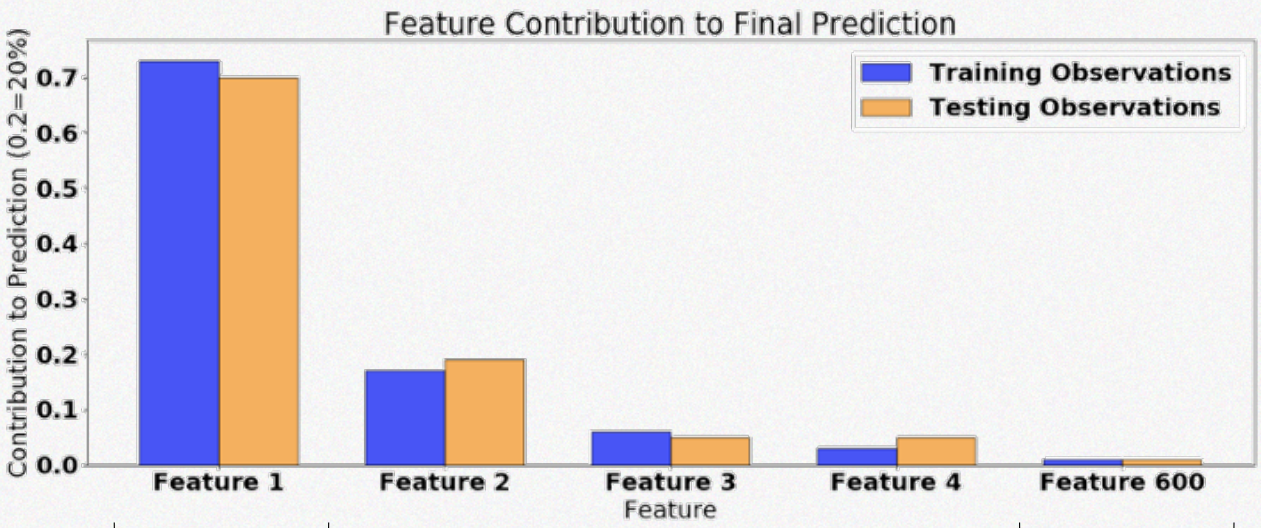
https://www.youtube.com/watch?v=ngOBhhINWb8

SHapley Additive exPlanations

The goal of SHAP is to explain the prediction of an instance x by computing the contribution of each feature to the prediction



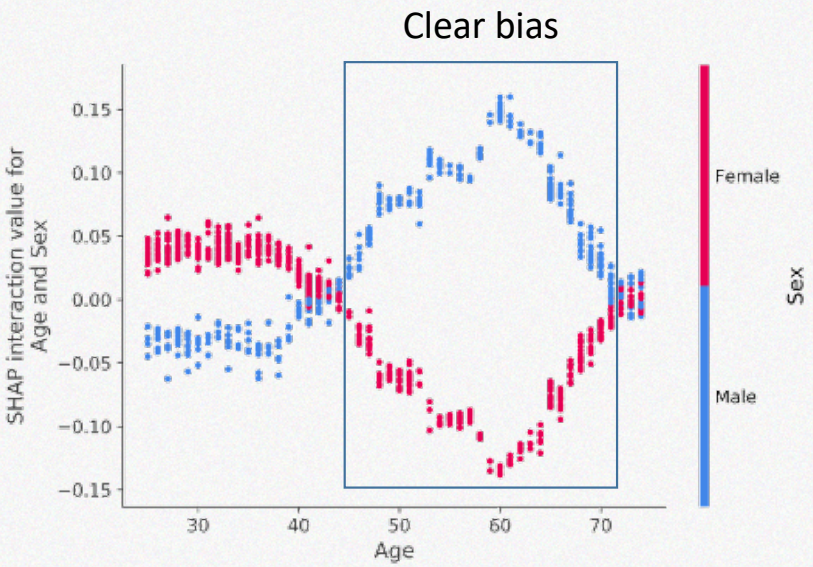
Feature Contribution

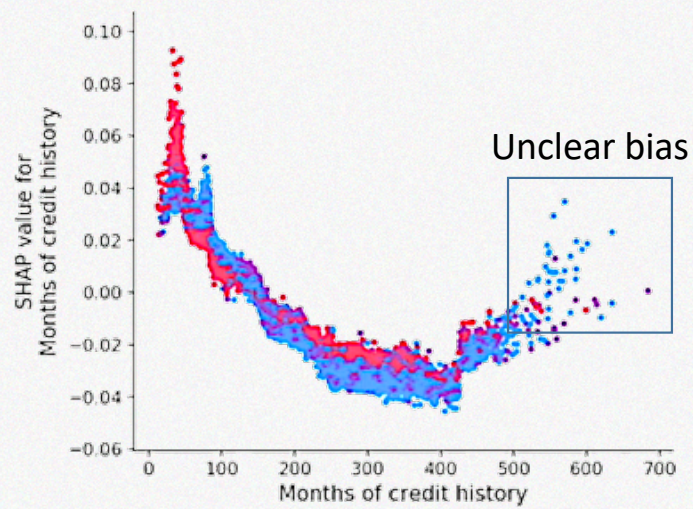


Marginal contribution of 70% of all predictions on average

Null average marginal contribution

How to Spot Biases using SHAP





3d) Evaluate model sensitivity

Sensitivity Assessment

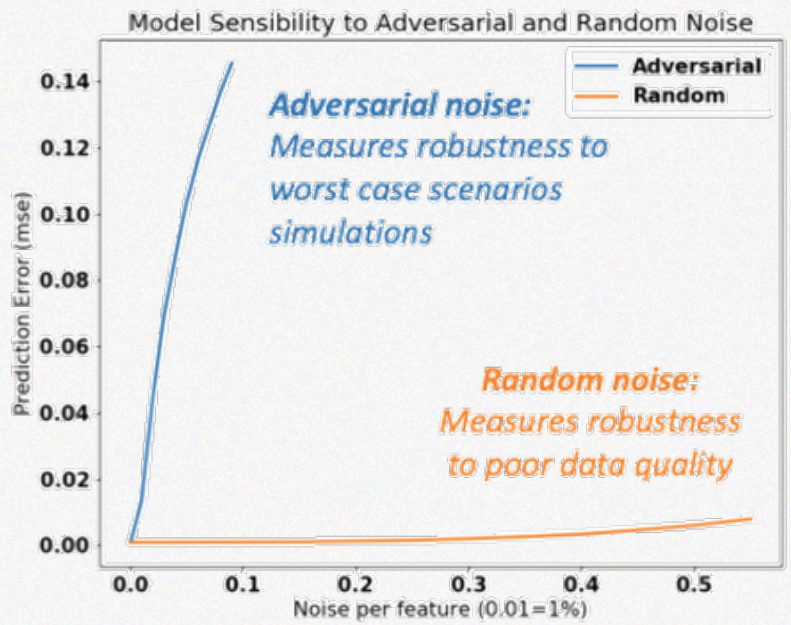
Noise is a slight modification in the input data

Random noise

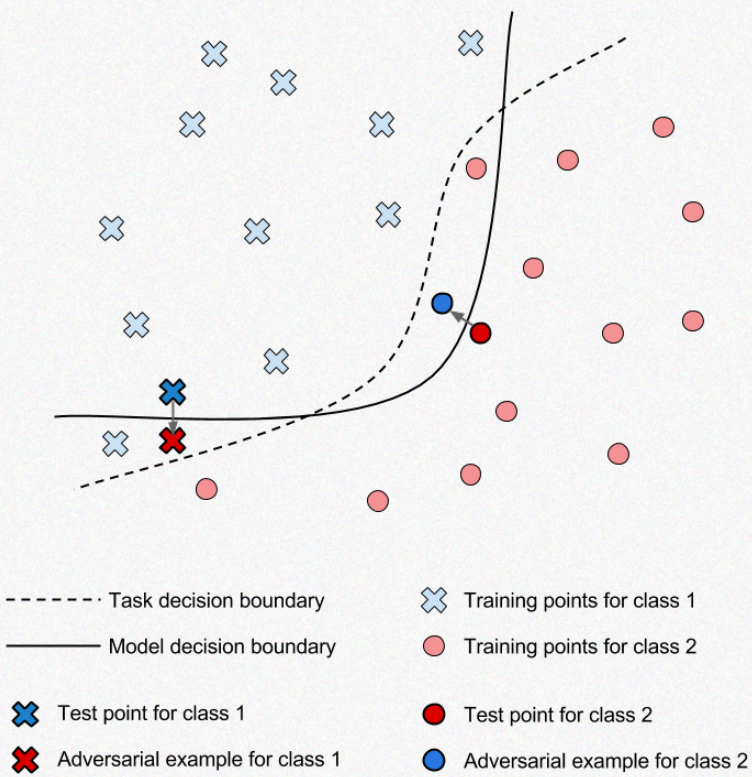
random noise can be added to each feature.

Adversarial noise

modifying each feature of the perfect amount and in the perfect direction to intentionally decrease the model's performance.



Adversarial Examples



$\boxed{X}$ $\boxed{\delta}$ $\boxed{X'}$

Original input + perturbation = adversarial example

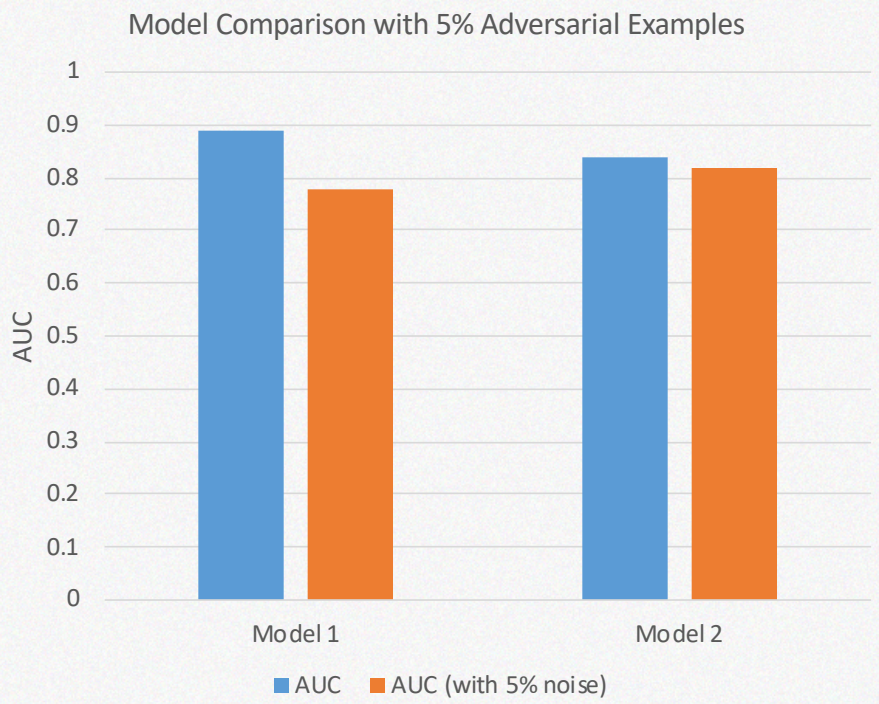
(correctly classified) (misclassified)

Adversarial examples differ based on the method used to craft the perturbation (e.g. Fast Gradient Sign Method, FGSM).

Adversarial Examples

Process

1. Build your models
2. Develop adversarial examples by model with the same level of noise
3. Assess dataset similarity (to be defined)
4. Compare models performance on adversarial examples
5. Assess most impacted features
6. Add most relevant examples to training dataset



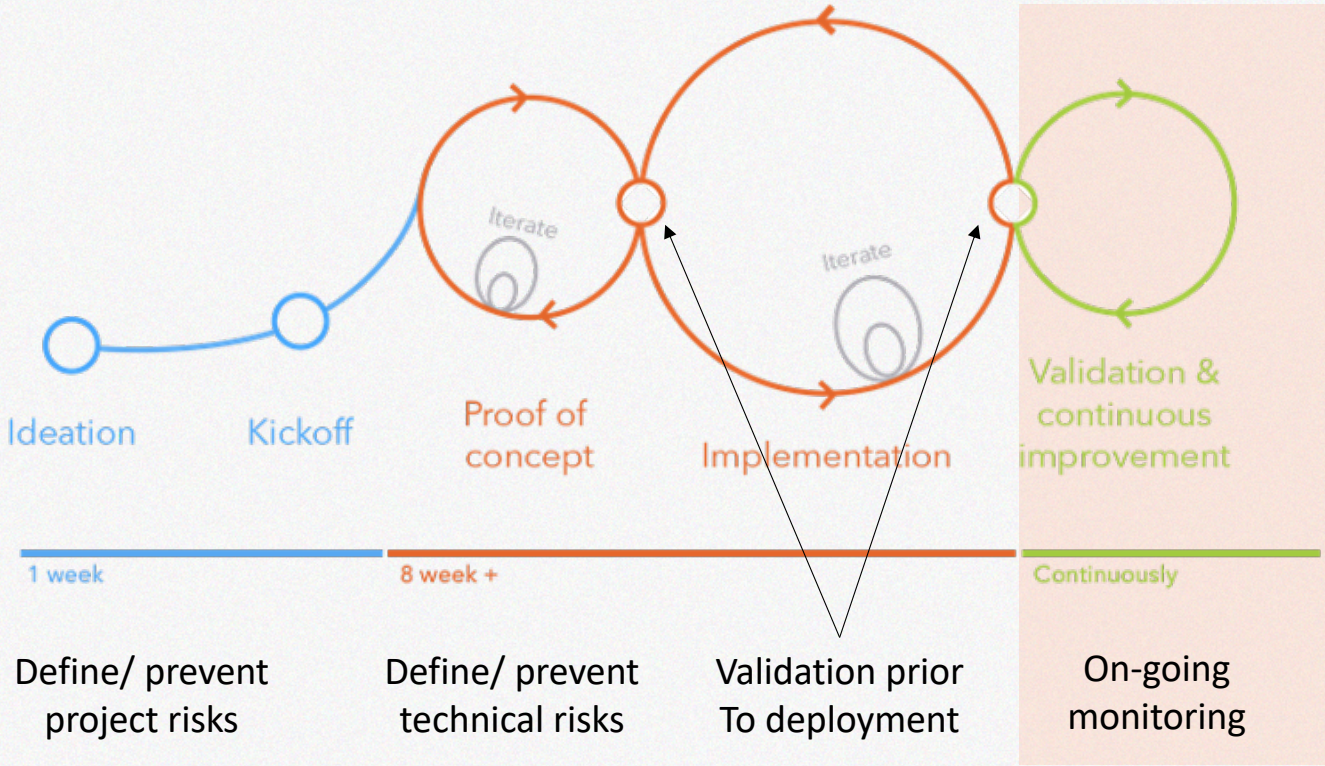
ODSC

MOOV AI

4. Monitoring Methodology

Initial Validation Process

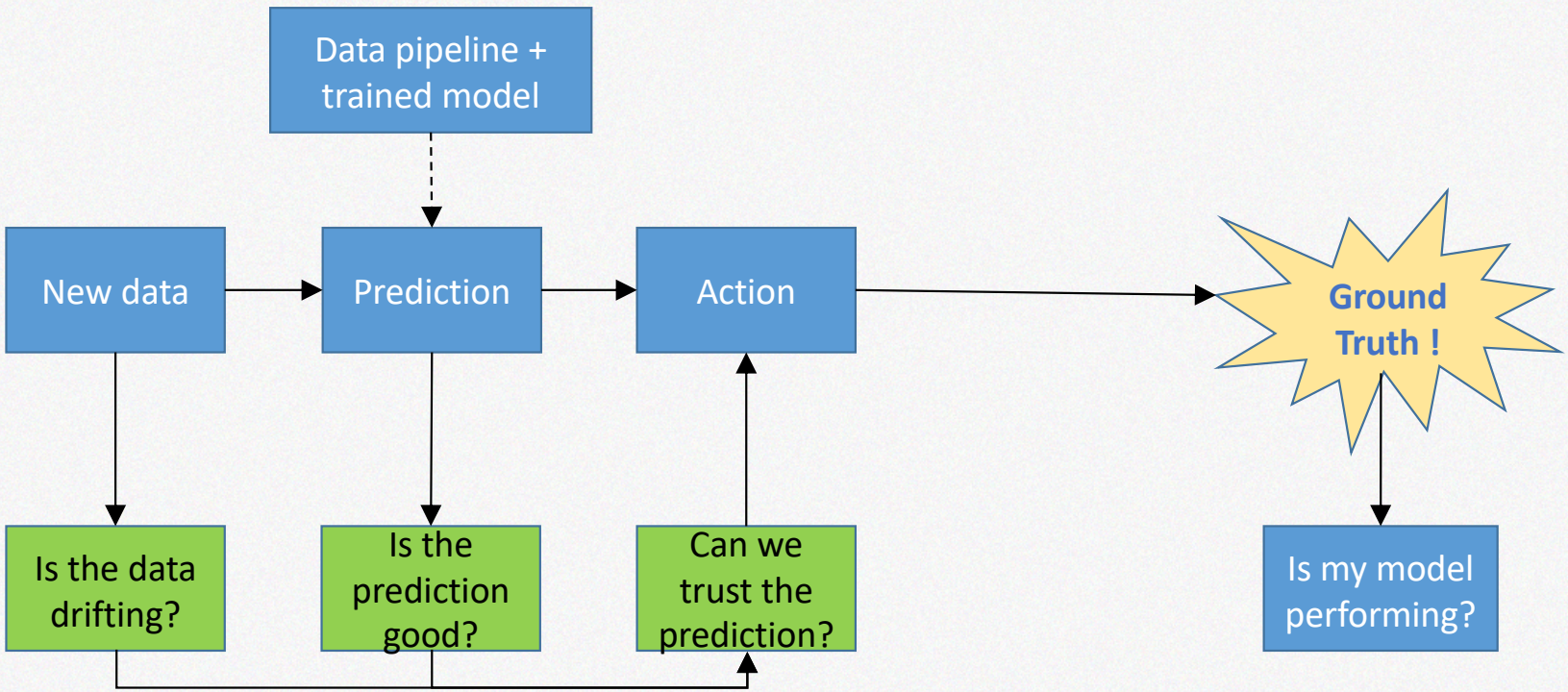
Project Development



Project Validation

Define/ prevent project risks	Define/ prevent technical risks	Validation prior To deployment	On-going monitoring

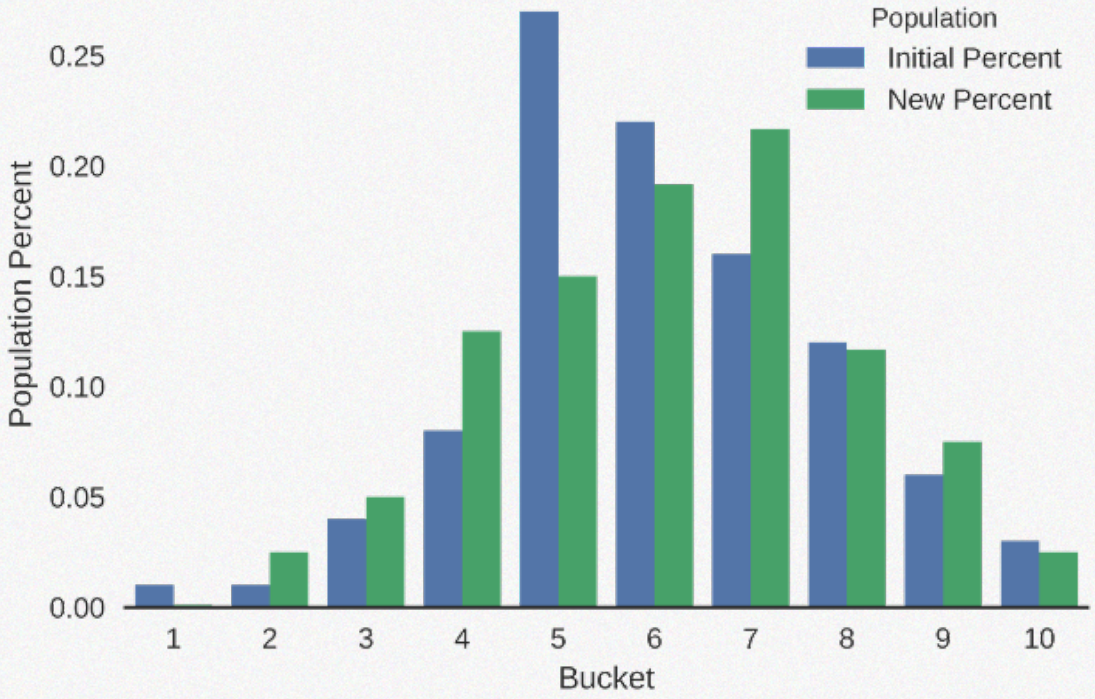
Monitoring Process



5. Monitoring tools

5a) Evaluate data drift

Population Stability Index



https://mwburke.github.io/data%20science/2018/04/29/population-stability-index.html

Population Stability Index

$$PSI = \sum \left(\left(Actual\% - Expected\% \right) \times ln\left(\frac{Actual\%}{Expected\%} \right) \right)$$

Rule of thumb

- **PSI < 0.1**: no significant population change
- **PSI < 0.2**: moderate population change
- **PSI >= 0.2**: significant population change

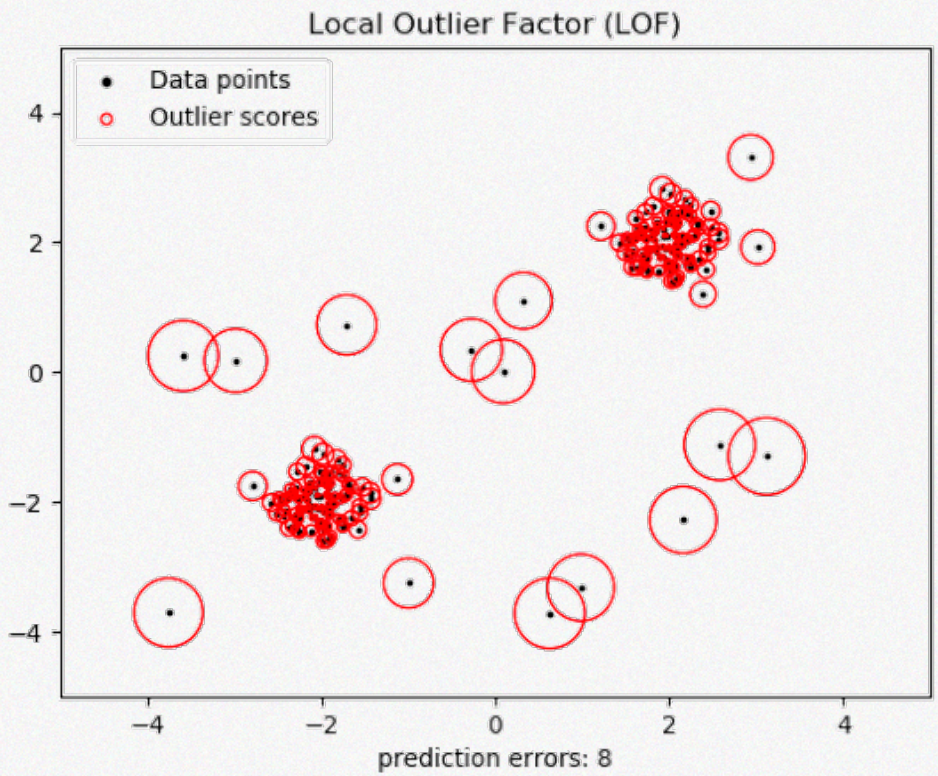
Pros

- Fast to run
- Easily interpretable
- No hyperparameters

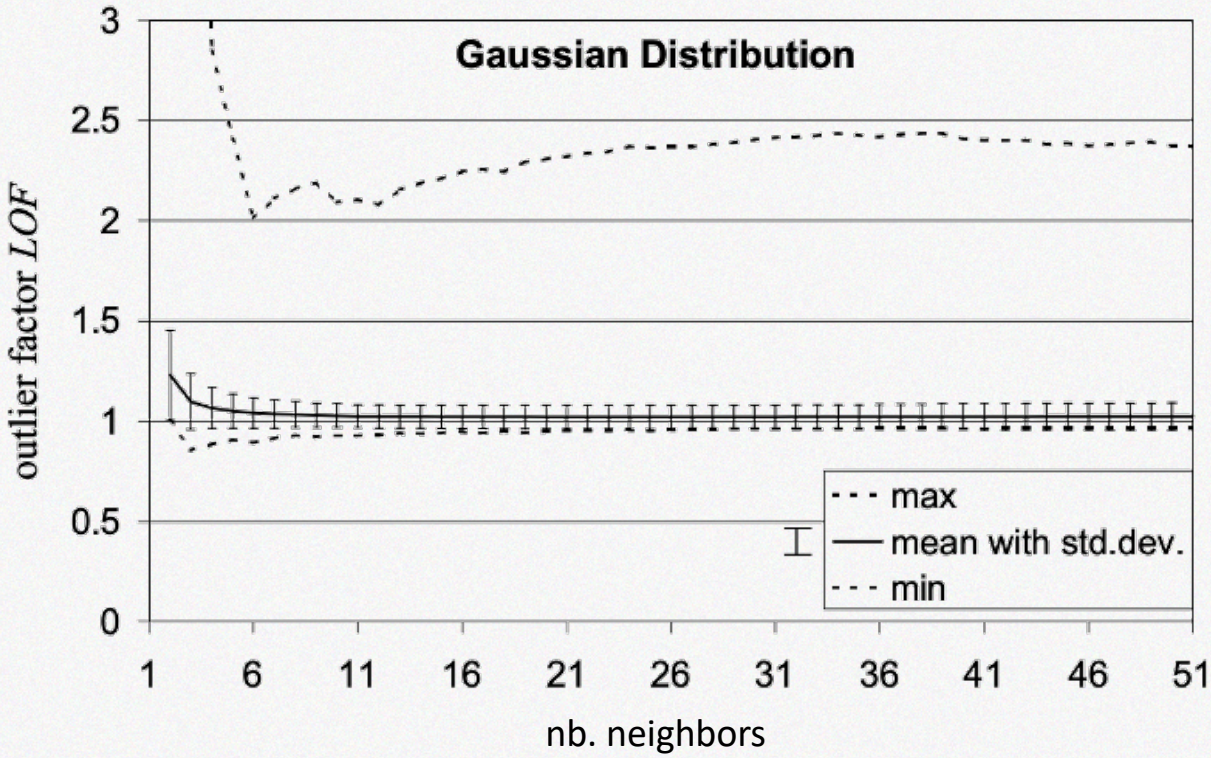
Cons

- Ignore feature interactions
- Runs on datasets only
- Not optimal for rare classes

Local Outlier Factor



Local Outlier Factor



Pros

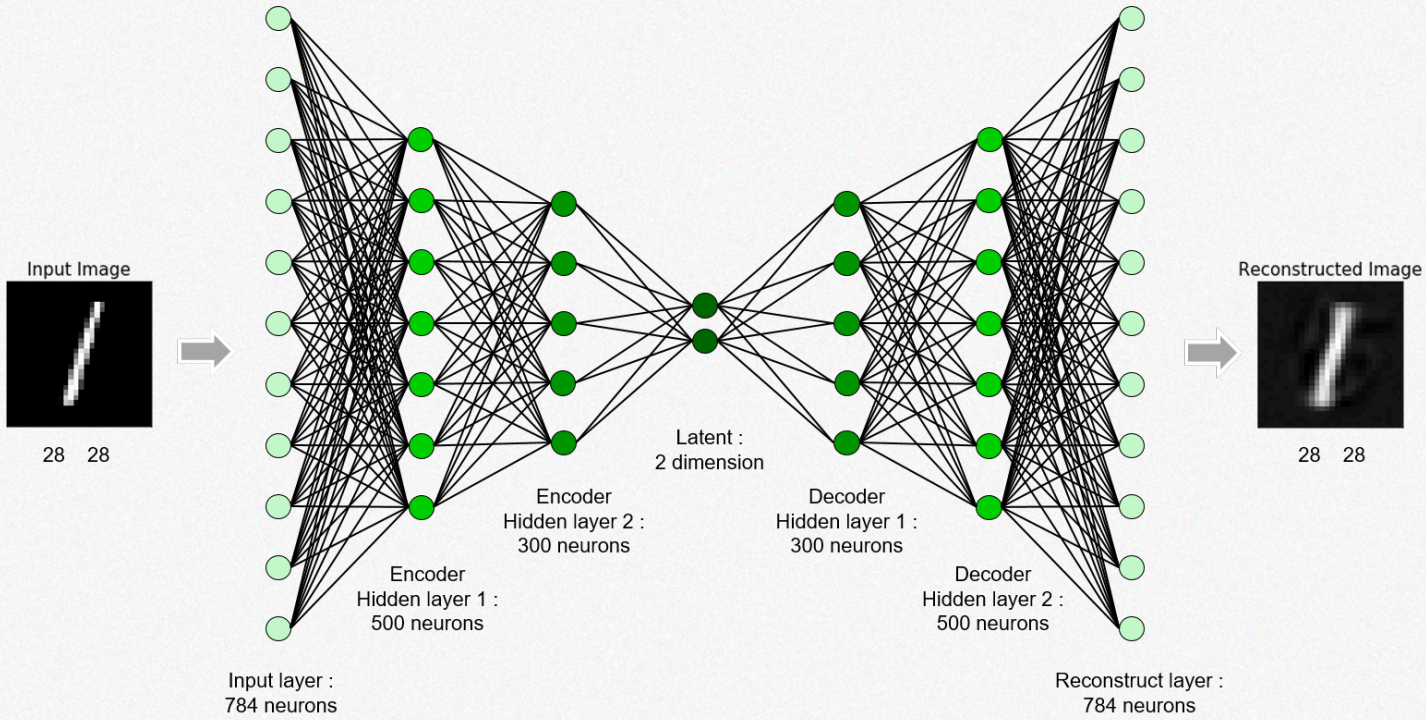
- Analyze relationship between features
- Few parameters to define
- Works with unique observation

Cons

- Slow to run
- Not ideal if too many dimensions
- Hard to interpret

https://www.dbs.ifi.lmu.de/Publikationen/Papers/LOF.pdf

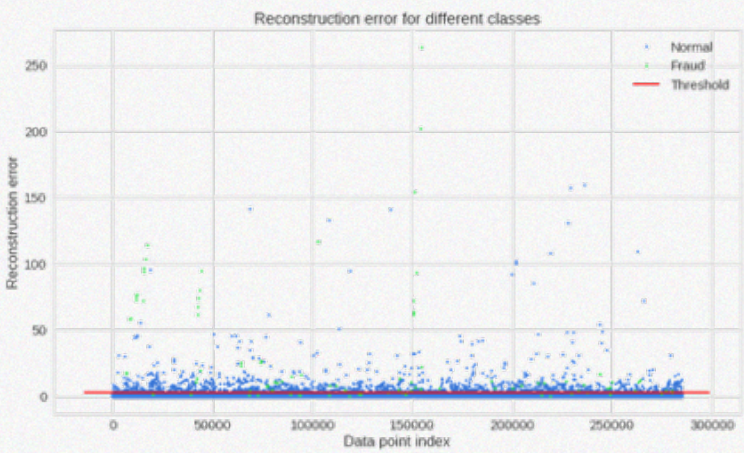
Autoencoder



Autoencoder

Data reconstruction error

$$L(x, x') = ||x - x'||^2$$



Pros

- Efficient on many dimensions
- Analyze relationship between features
- Works with unique observation

Cons

- Hard to interpret
- Complex to configure
- Requires a big dataset

5b) Evaluate model degradation

Monitor Prediction Performance

Prediction degradation can be identified by an augmentation of uncertainty

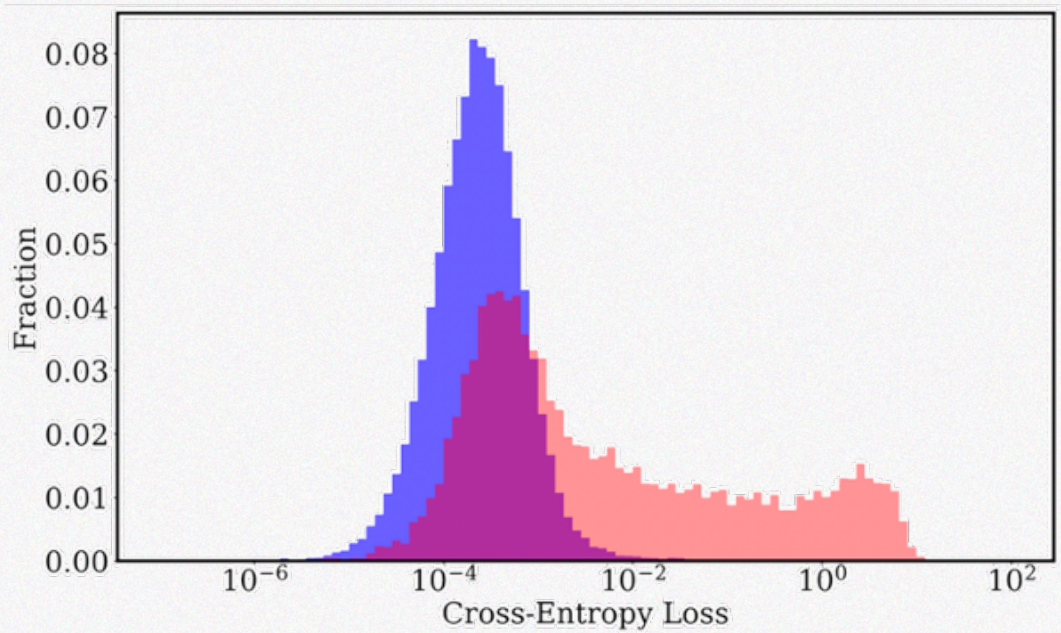
For classification

- a smaller ratio or predictions near 0 and 1

For regressions

- a bigger confidence interval

Cross-Entropy on New Predictions

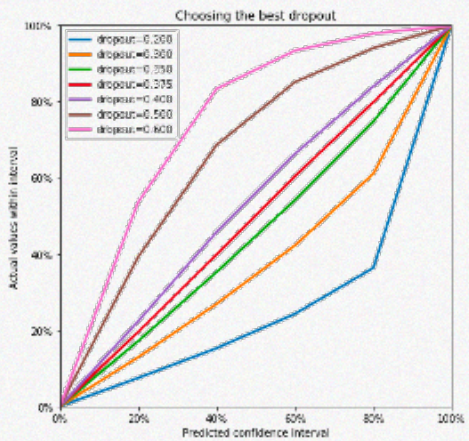


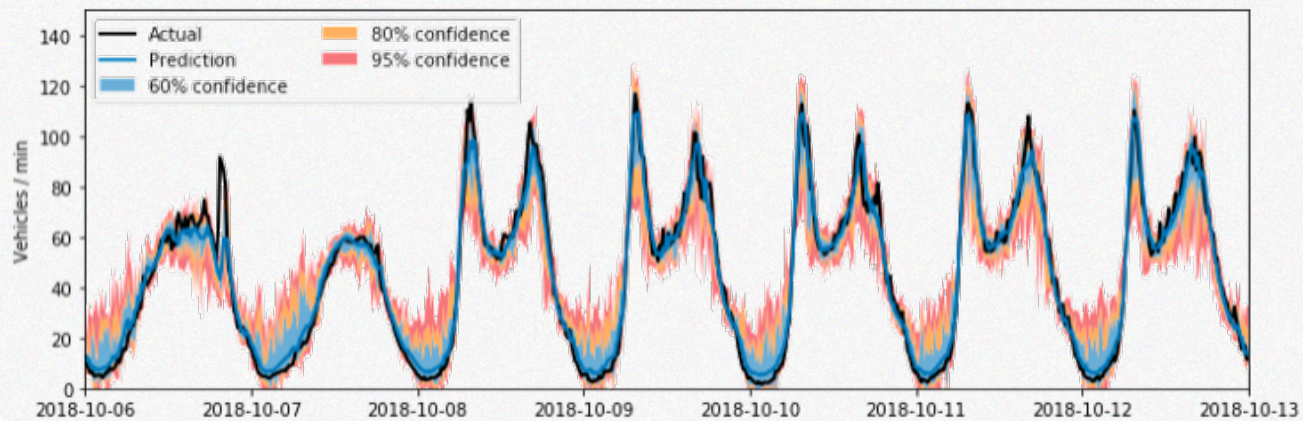
Use the predicted class as truth to estimate the prediction degradation

Again, rule of thumb
Cross-Entropy = 0.0: Perfect probabilities
Cross-Entropy < 0.1: moderate uncertainty
Cross-Entropy > 0.3: Great uncertainty

https://www.groundai.com/project/simultaneous-classification-and-novelty-detection-using-deep-neural-networks/1

Confidence Interval in Deep Learning

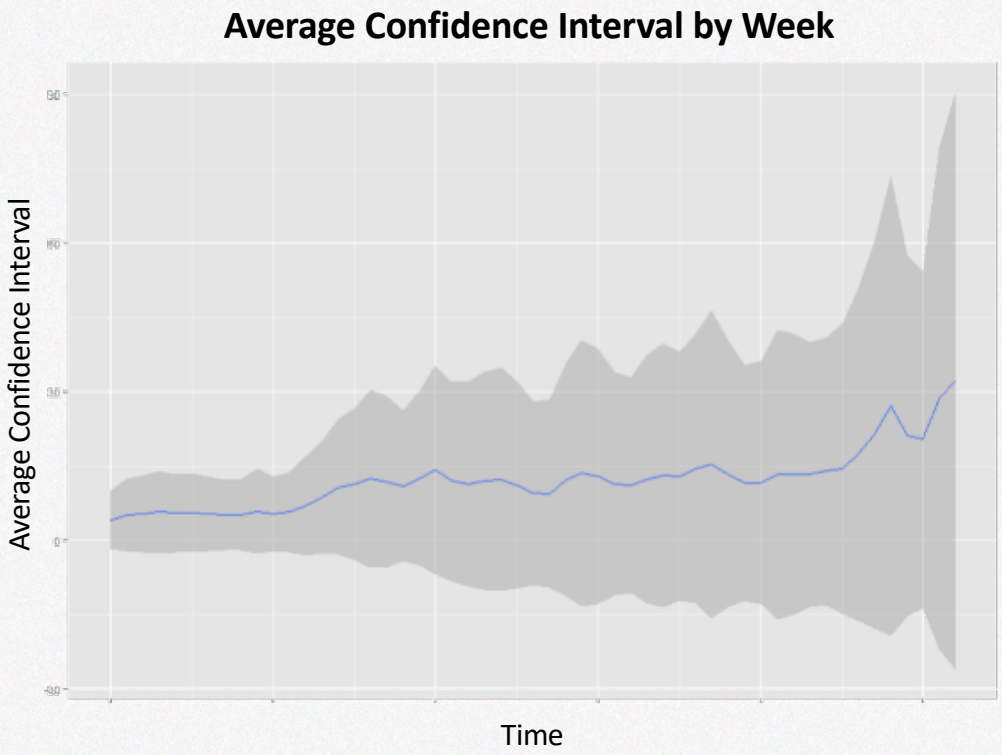




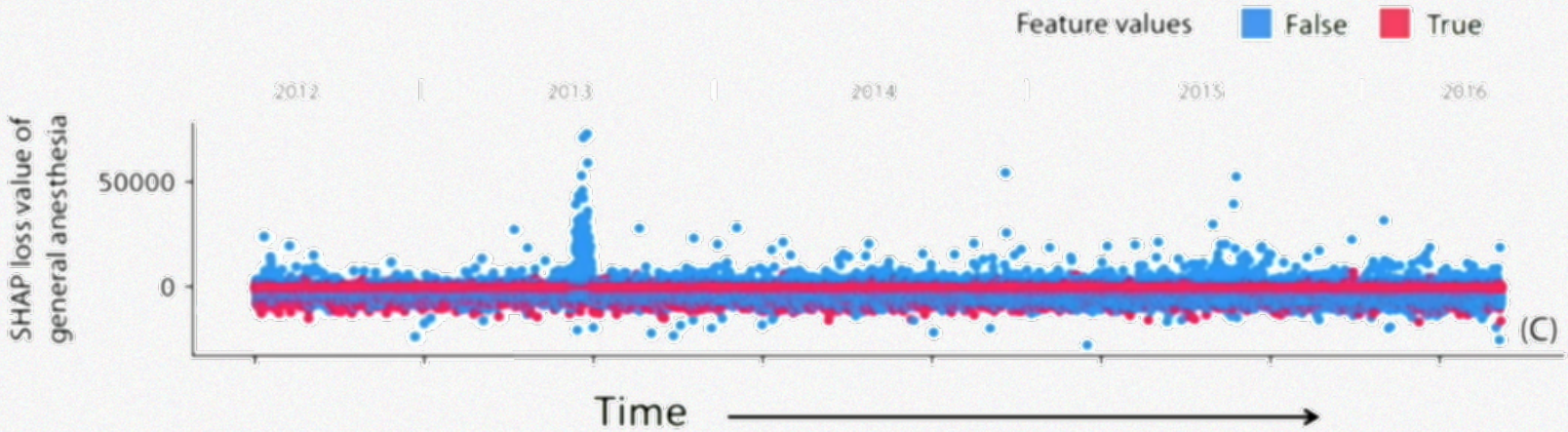
Using dropout during prediction allows the creation of a range of possible values!!!

https://medium.com/hal24k-techblog/how-to-generate-neural-network-confidence-intervals-with-keras-e4c0b78ebbdf

Confidence Interval

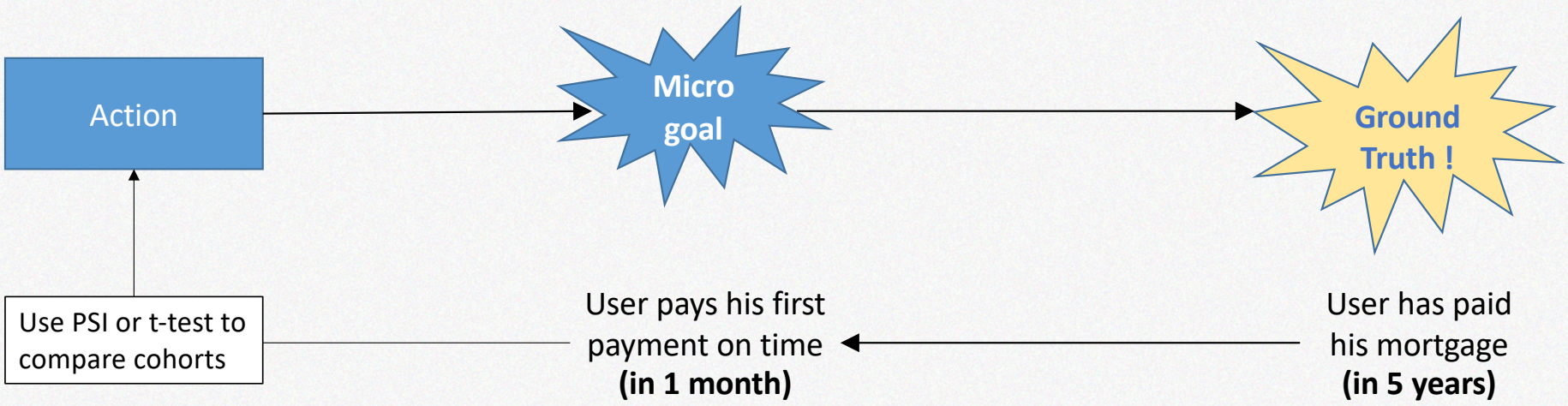


SHAP loss value



5c) Evaluate label stability

Label Stability



Questions ?

MOOV AI

Thank you

Olivier Blais
Olivier.blais@moov.ai