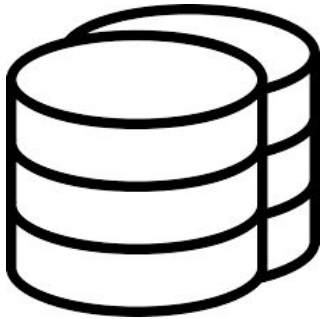
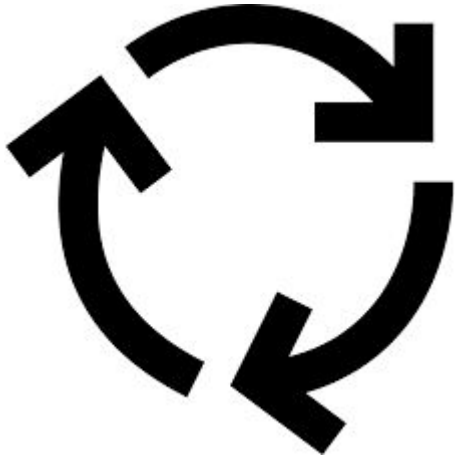
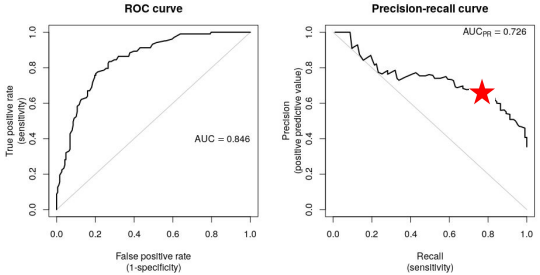
Model Performance Management

Dan Salo
December 7th, 2022

ML From Scratch







Training and Validation Data Cleaning

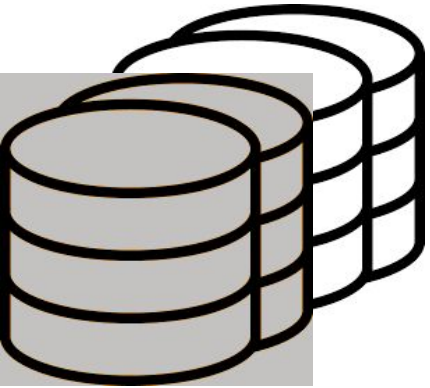
Model Training and Hyperparameter Tuning

Model Evaluation and Threshold Selection

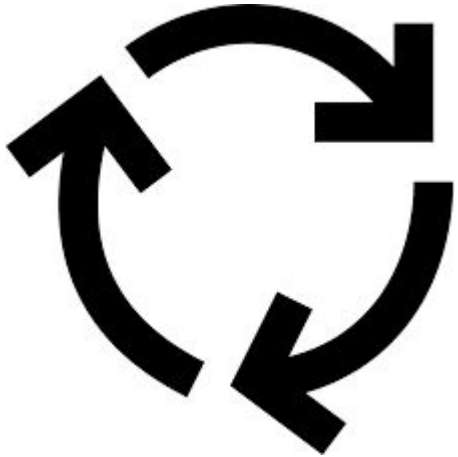
But then Wrong Answers …

Source: xkcd

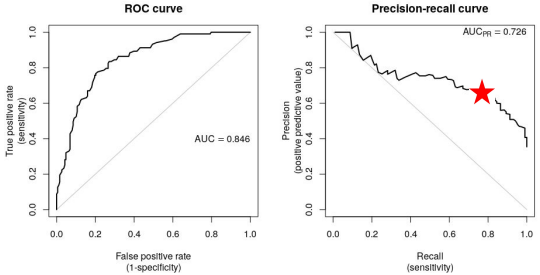
Naive Retraining



Training and Validation
Data Cleaning
with New Labeled Data



Model Training and
Hyperparameter
Tuning



Model Evaluation and
Threshold Selection

Outline

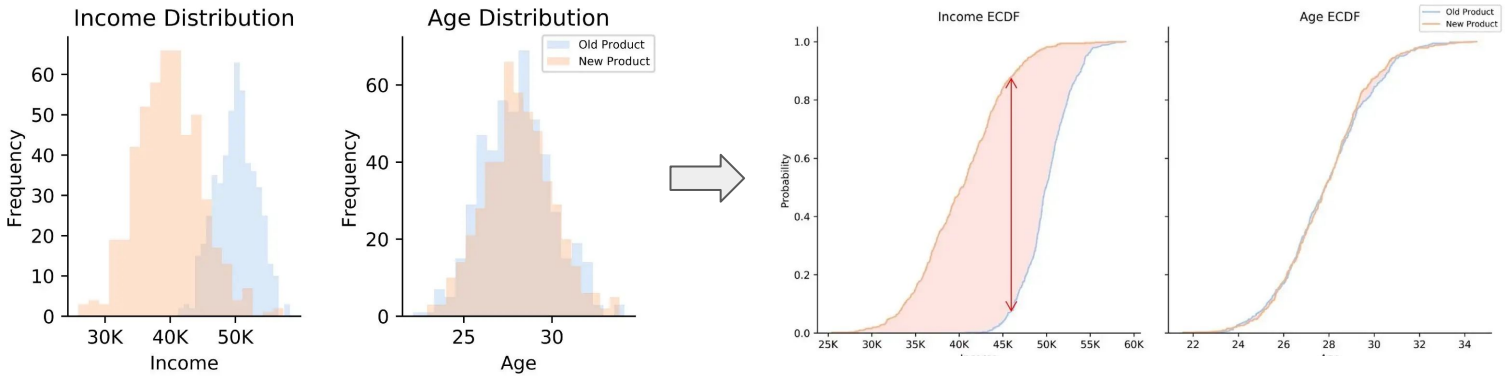
- When to retrain an ML model?
- Which inference data to label for retrainings?
- How to prevent regression errors between retrainings?

When to Retrain an ML Model?

Drift Detection

Data Drift: Univariate Features in Training and Test Sets

Continuous	**Discrete**
Kolmogorov-Smirnov Jensen-Shannon Wasserstein	Chi^2 Jensen-Shannon Population Stability Index

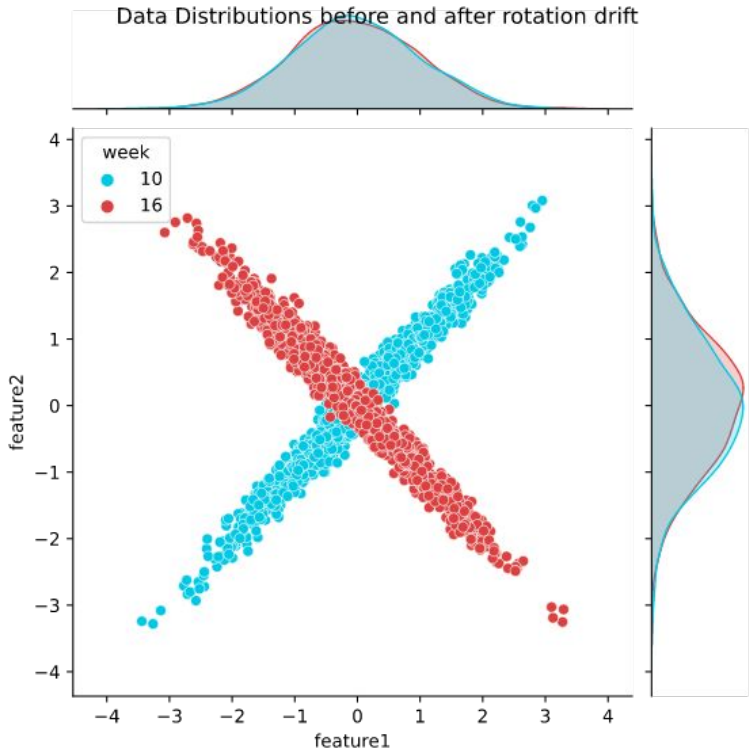


Source: Blog

Requires *training* data at *inference* time.

Data Drift: Multivariate Features in Training and Test Sets

1. Fit PCA on training data. Save Transform Model.
2. Apply Transform to inference data to generated Components.
3. Reconstruct inference data using Components.
4. Alert if reconstruction error crosses threshold.



Source: NannyML

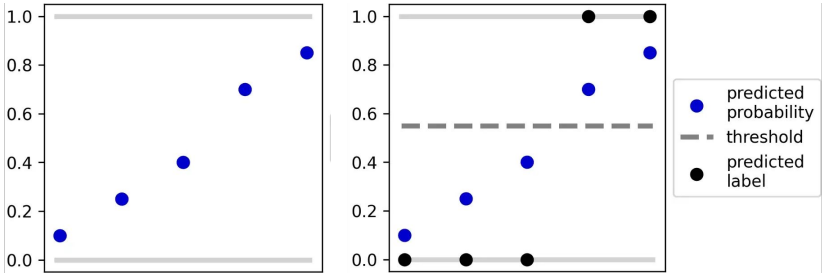
Requires *Transform Model* at *inference* time.

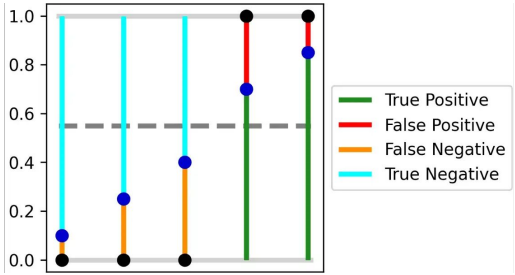
Model Drift

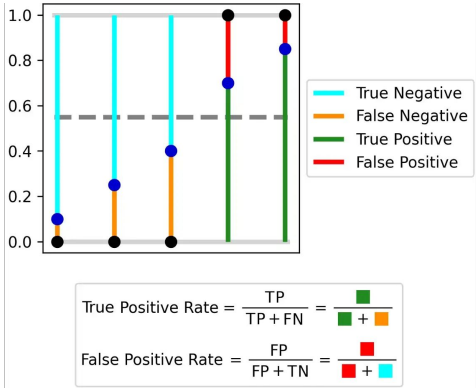
If Precision / Recall on *labeled* test set is very different from training set, it's time to retrain.

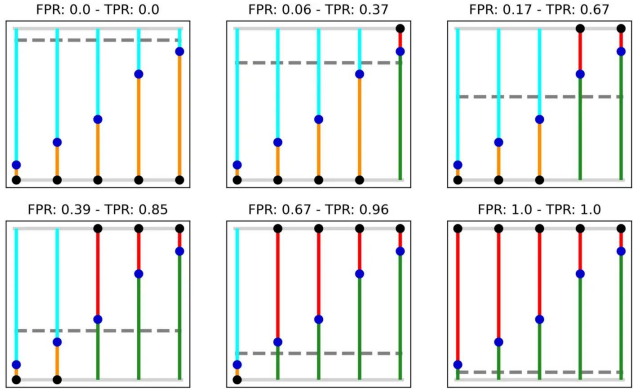
What if there was a way to calculate Precision / Recall / etc *without labeled data*?

Model Drift: Confidence-Based Performance Estimation









Calculating Precision and Recall …

Calculating ROC Curve …

Source: Blog

Requires *probabilities on inference* at inference time.

But my scikit-learn model yields probabilities …

predict_proba(*X*)

Probability estimates.

The returned estimates for all classes are ordered by the label of classes.

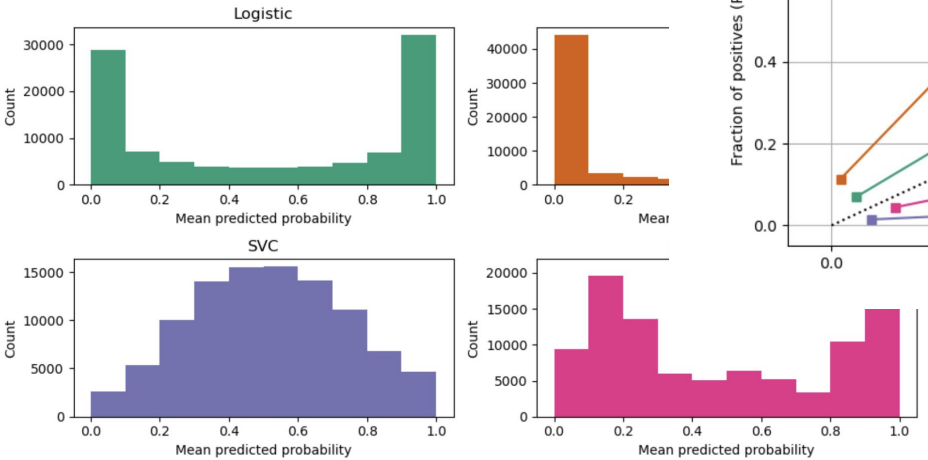
For a multi_class problem, if multi_class is set to be "multinomial" the softma

Model confidence scores do sum to 1… but that doesn't make them valid probas.

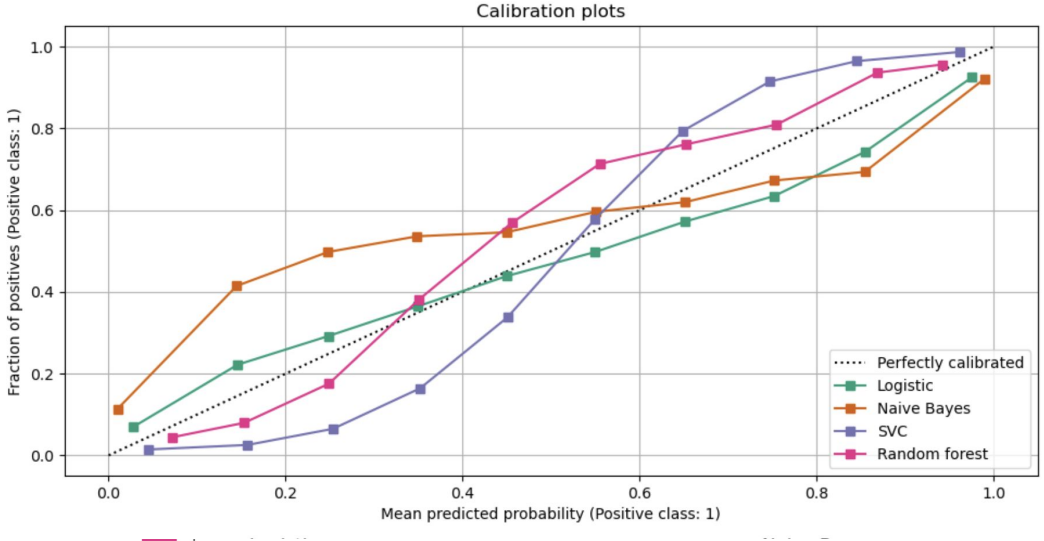
We want to be able to say: "5% of all examples with score of ~0.05 are true positives"

Source: Blog

Model Calibration



Characteristic Output of ML Models



Calibrated Model Outputs

sklearn.calibration.CalibratedClassifierCV

class sklearn.calibration.**CalibratedClassifierCV**(*base_estimator=None, *, method='sigmoid', cv=None, n_jobs=None, ensemble=True*) [source]

Source: Scikit-Learn Docs

Requires *training* data at *train* time.

Beyond Calibration: Conformal Predictions

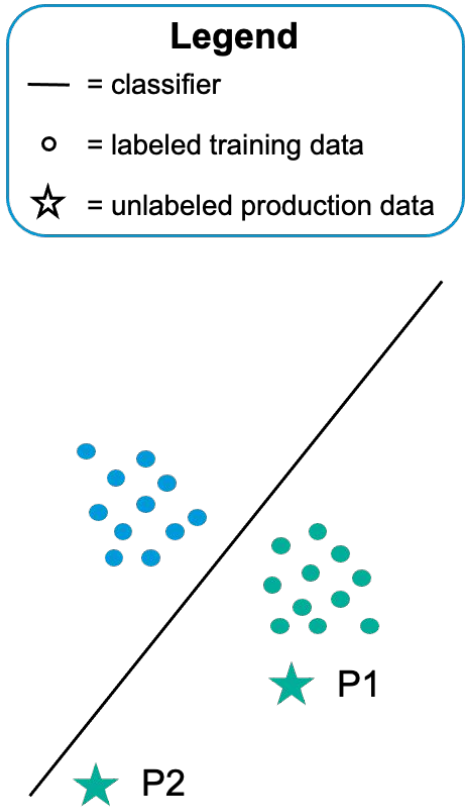
Ask Brian or WSU

Source: Brian's Brain

Requires *training* data at *inference time*?

Beyond Calibration: Trust Score

- Trust score = $\frac{\text{Distance to closest } \textit{non-predicted} \text{ label group}}{\text{Distance to } \textit{predicted} \text{ label group}}$
- **High trust:** Data is close to its predicted label group vs. other label groups
  - P1: Close to green (predicted) group vs. blue group
- **Low trust:** Data is as close to non-predicted label groups as predicted label
  - P2: (Almost) equally close to predicted group vs. other group



Source: Trust Scores

Requires *training* data at *inference time*.

Platform Requirements Summary

Method	**Requirements**
Univariate Feature Diff	Training Set at Inference Time
	Unlabeled Inference Sample at Inference Time
Multivariate Feature Diff	Transform Model at Inference Time
	Unlabeled Inference Sample at Inference Time
CBPE	Training Set at Inference Time
	Inference Sample with Predictions at Inference Time

Which data to label?

Active Learning

Active Learning Ideas

Group similar examples based on Approximate Nearest Neighbor techniques and select one exemplar from each group.

Use outlier detection to select rare examples based on feature distributions of training set.

Use model output to select examples with low confidence / probability / large distance from training set.

Requires *predictions on inference.*

Platform Requirements Summary

Method	**Requirements**
Clustering	ANN Algorithm
	Inference Sample
Outlier Detection	Outlier Algorithm
	Inference Sample
Ranking with Model Output	Transform Model
	Inference Sample with Predictions

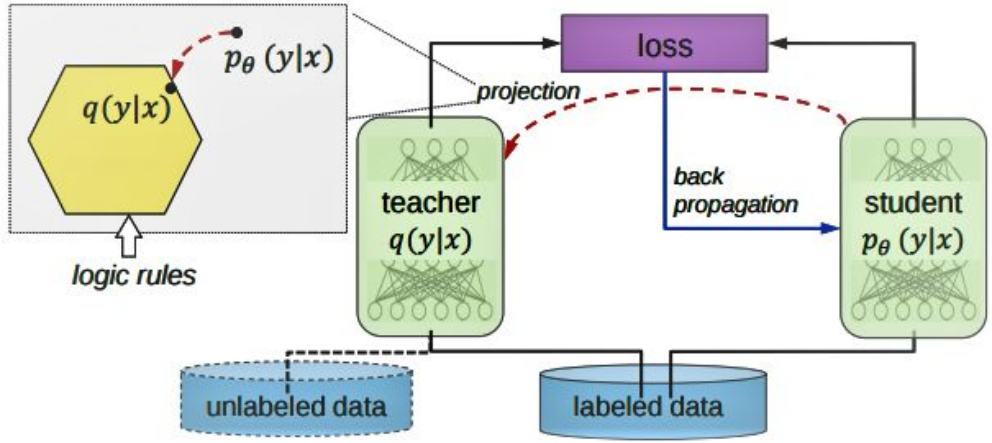
How To Prevent Regression Errors?

Prediction Churn

Knowledge Distillation for Mitigating Prediction Churn

Ask Zack / Ryan for MIDS Capstone Project Slides / Demo Link.

TLDR: incorporate the logits of the previous model (teacher) when that model had the correct prediction when training the new model (student).



Source: Blog

Requires previous model available at train time.

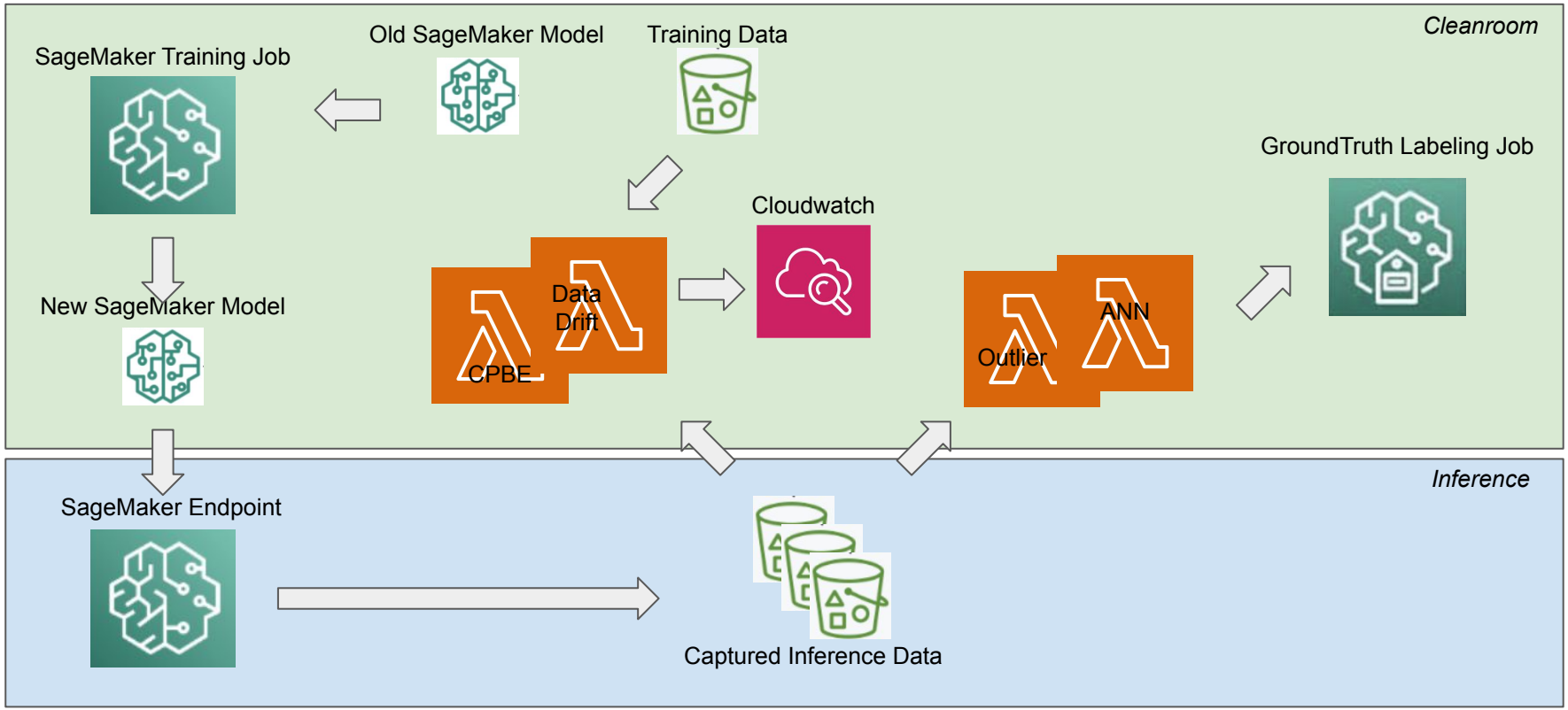
ML Platform

Where Dreams Become Reality™

Wishlist for Drift Detection and Active Learning

- Mechanism to capture inference data with model predictions / probabilities
- Exposing training data with model predictions / probabilities to other algorithms
- Exposing old model to new model training
- Ad-hoc algorithm job runner on datasets
- Labeling inference for data

Architecture Diagram



Thank You

Questions?

References

Matchmaker: Data Drift Mitigation (2022)

Overview of Unsupervised Drift Detection (2020)

Characterizing Concept Drift (2013)

To Trust or Not To Trust a Classifier (2020)

Awesome Conformal Prediction (2022)

References

NannyML Docs

Seldon: Monitoring and Explainability of Models in Production

AWS: Detect Data Drift in Production

Scikit-Learn: Probability Calibration