Data Challenge 4

Model Drift Management

Matt Mitchell | *Sumit Dhar*

Agenda

Day-1

Session 1: Concepts & Discussions

Introducing Model Drift
Types of Drift
- Data Drift
- Concept Drift

Key Quality Metrics
- Population Metrics
- Model Performance Metrics

Revisiting the Engine Failure Data
Illustrative toy model: Dead in 50 cycles
- Interpreting population & model metrics

DataBricks: Production Model Deployment

Day-1

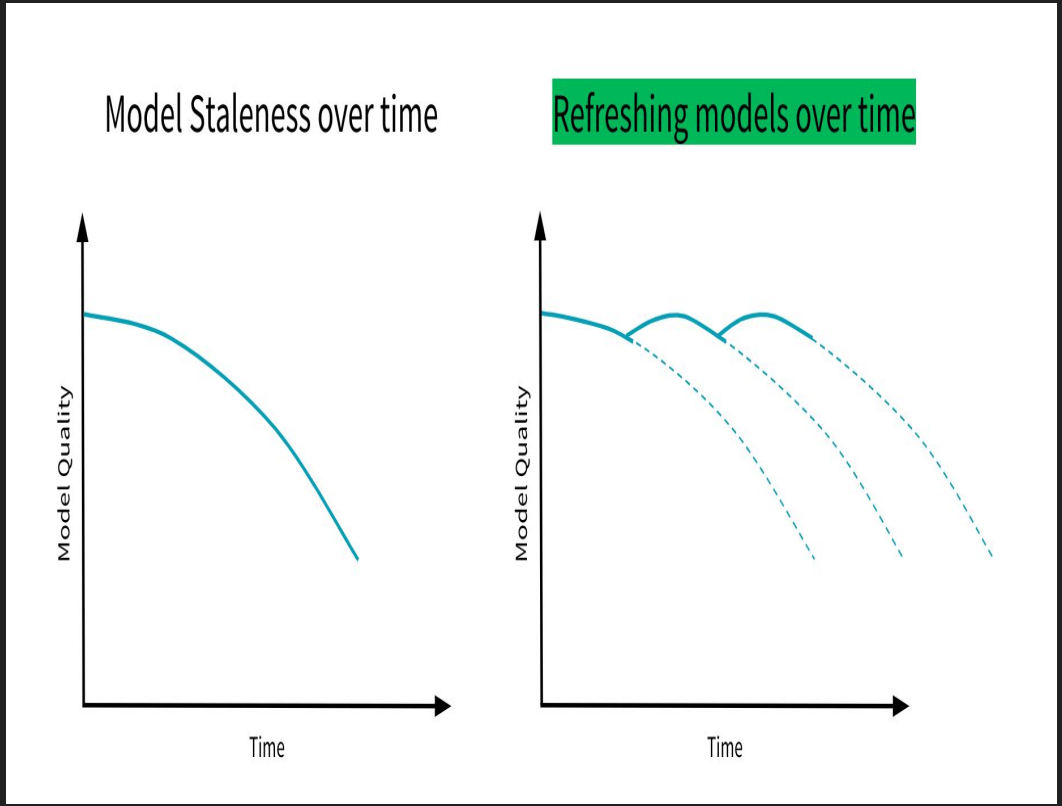
Session 2: Industrial Strength Model Drift Management

- Model Management: Component Technologies
- Hands Down: Apply Model Drift concepts to RUL Data

Session- 1

Introducing the Concept of Drift

- Production deployment of ML models is the beginning of the life cycle of model management
- The deterioration of predictive power of a model is called *Model Drift*
- In order to ensure continued performance ML models need to be monitored and refreshed from time to time



Major Types of Drift

Model drift can occur when there is some form of change to feature data or target dependencies.

Data Drift

The features used to train a model are selected from the input data. When statistical properties of this input data change, it will have a downstream impact on the model's quality.

E.g Time Varying Change in Data

Concept Drift

When statistical properties of the target variable change, the very concept of what you are trying to predict changes

E.g Definition of Failure Mode or Failure Reason

Protecting Against Drift

Instrument feedback loop from a monitoring system, and refreshing models over time, will help avoid model staleness.

Scenarios where you can deploy monitoring:

- Training data
- Schema & distribution of incoming data
- Distribution of labels

Requests & predictions

- Schema & distribution of requests
- Distribution of predictions
- Quality of predictions

Examining Drift: Statistics & Performance Metrics

Population/ Sample Statistics

Population Stability Index (PSI): Measure of population stability between two samples

Kolmogorov-Smirnov test (or KS test) is a nonparametric test of the equality of distributions - - compare two samples (two-sample K–S test).

Kullback–Leibler (or KL) divergence is a measure of difference between two distributions

Jensen-Shannon (or JS) divergence is a method of measuring the similarity between two distributions.

Model Performance Statistics

Classification Efficiencies- best value at 1 and worst score at 0

Precision: The ability of the classifier not to label as positive a sample that is negative

Recall: the ability of the classifier to find all the positive samples.

F1 Score: Weighted average of the precision and recall, where an F1 score reaches its

ROC- Area under the ROC Curve [> .5]

Population Stability Index (PSI)

PSI is a measure of how much a population has shifted between two different samples of a population.

It does this by bucketing the two distributions and comparing the percents of items in each of the buckets

The common interpretations of the PSI result are:

- PSI < 0.1: no significant population change
- PSI [0.1, 0.2]: moderate population change
- PSI >= 0.2: significant population change

Good decisions in a machine learning model building context

- Monitoring the PSI score from the time of model training to current time can be used as **automatic triggers** to re-train the model when PSI passes a certain threshold
- Exclude feature prone to changes in distribution from ML Models

PSI is an easy way to check the volatility of population changes of features by comparing several previous time periods populations

KS Statistics

This is a two-sided test for the null hypothesis that 2 independent samples are drawn from the same continuous distribution.

If the KS statistic is small or the p-value is high [p > 1%], then we cannot reject the hypothesis that the distributions of the two samples are the same.

KS is non parametric and based on a significance test

Lot more statistically solid

Used for:

- Proactive feature selection
- Reactive Model Management

Classification Quality Metrics- Definition

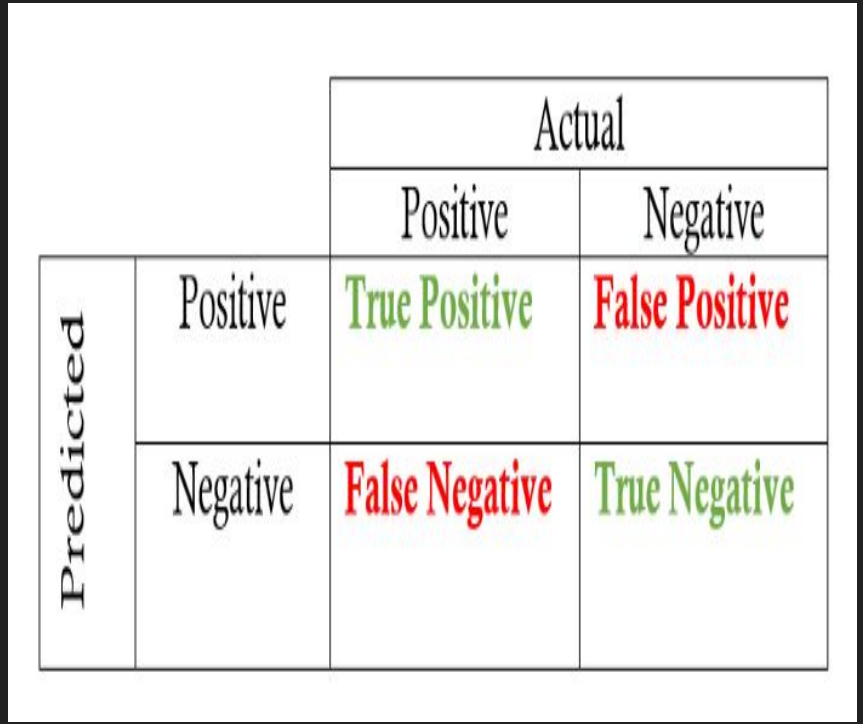
Precision = (TP)/(TP+FP)

Recall = (TP)/(TP+FN)

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

AOC:- Area under the ROC curve

.5 in case of a binary random selection



Revisiting the C-MAPPS Data: All the Data Sets

0	1	2	3
Data Set: FD001 Train trajectories: 100 Test trajectories: 100 Conditions: ONE (Sea Level) Fault Modes: ONE (HPC Degradation)	Data Set: FD002 Train trajectories: 260 Test trajectories: 259 Conditions: SIX Fault Modes: ONE (HPC Degradation)	Data Set: FD003 Train trajectories: 100 Test trajectories: 100 Conditions: ONE (Sea Level) Fault Modes: TWO (HPC Degradation, Fan Degradation)	Data Set: FD004 Train trajectories: 248 Test trajectories: 249 Conditions: SIX Fault Modes: TWO (HPC Degradation, ***Fan Degradation***)

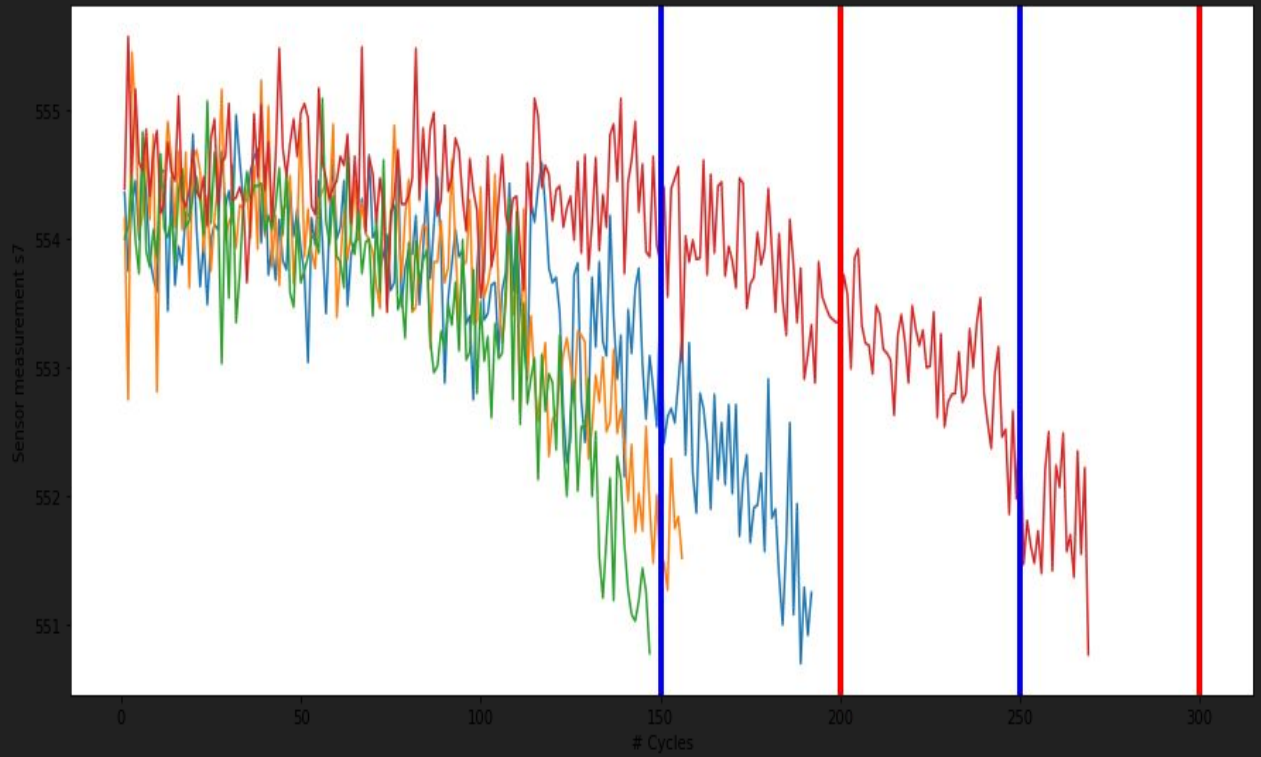
Toy Model to illustrate Effect: “Dead in 50 Cycles”

A simple logistics regression model to predict engine failure in the next 50 cycles

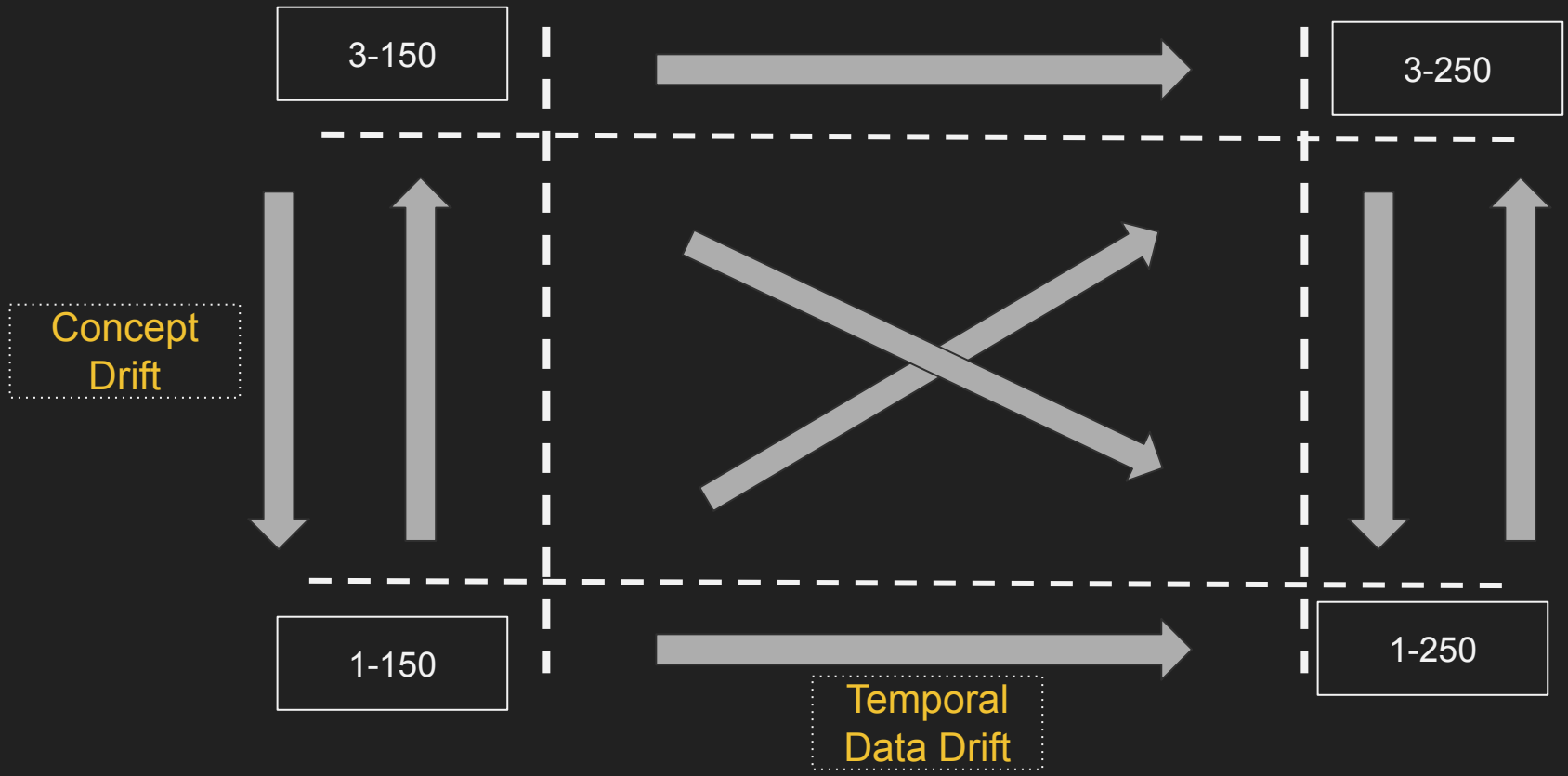
```
LogisticRegression(random_state=42,solver='saga', max_iter=1000, penalty='l1').fit(X, y)
```

X -> Averaged Sensor Measurements in past 50 cycles

y -> Binary class[1,0]



Model Topology: Simulating Drifts



Model Performance Metrics Discussion

	1-150	1-250	3-150	3-250
1-150	**F1: 0.75 Precision: 0.75 Recall: 0.75 ROC: 0.77**	F1: 0.89 Precision: 0.88 Recall: 0.91 ROC: 0.79	F1: 0.75 Precision: 0.85 Recall: 0.66 ROC: 0.81	F1: 0.57 Precision: 0.88 Recall: 0.43 ROC: 0.68
1-250	**X**	**F1: 0.84 Precision: 0.73 Recall: 1.0 ROC: 0.5**	**X**	F1: 0.68 Precision: 0.51 Recall: 1.0 ROC: 0.5
3-150	F1: 0.7 Precision: 0.74 Recall: 0.66 ROC: 0.73	F1: 0.89 Precision: 0.88 Recall: 0.91 ROC: 0.79	**F1: 0.75 Precision: 0.85 Recall: 0.66 ROC: 0.81**	F1: 0.59 Precision: 0.89 Recall: 0.44 ROC: 0.69
3-250	**X**	F1: 0.79 Precision: 0.81 Recall: 0.78 ROC: 0.64	**X**	**F1: 0.83 Precision: 0.88 Recall: 0.8 ROC: 0.84**

KS Statistics: Null Hypothesis Rejects

	1-150	1-250	3-150	3-250
1-150	X	's11', 's13', 's14', 's16'	's11', 's15'	's14', 's15'
1-250	X	X	X	's4', 's15'
3-150	's11', 's15'	's3', 's4', 's9', 's11', 's14', 's16', 's17'	X	's14', 's15'
3-250	X	's4', 's15'	X	X

PSI: 5 Highest PSIs

	1-150	1-250	3-150	3-250
1-150	X	's6', 's20', 's14', 'op_setting2', 's17'	'op_setting1', 's20', 's18', 'op_setting2', 's7'	's12', 's18', 's17', 's16', 'op_setting2'
1-250	X	X	X	'op_setting1', 's12', 's18', 's5', 's16'
3-150	'op_setting3', 'op_setting1', 's7', 's9', 's20'	's2', 's14', 's9', 's6', 's17'	X	'op_setting1', 's16', 's9', 's6', 's18'
3-250	X	's4', 's13', 's11', 's14', 's7'	X	X

Addressing Drift

- Develop a single “best” model once & operationally manage future scoring
- First-level intervention:
  - Periodically update your static model with more recent historical data
  - Use a sliding window
- Update using existing model
- Ensemble approach: New model learns to correct the predictions from the static model based on the relationships in more recent data
- Instrument to detect changes and choose a specific and different model to make predictions
- Make model robust through data preparation

Addendum