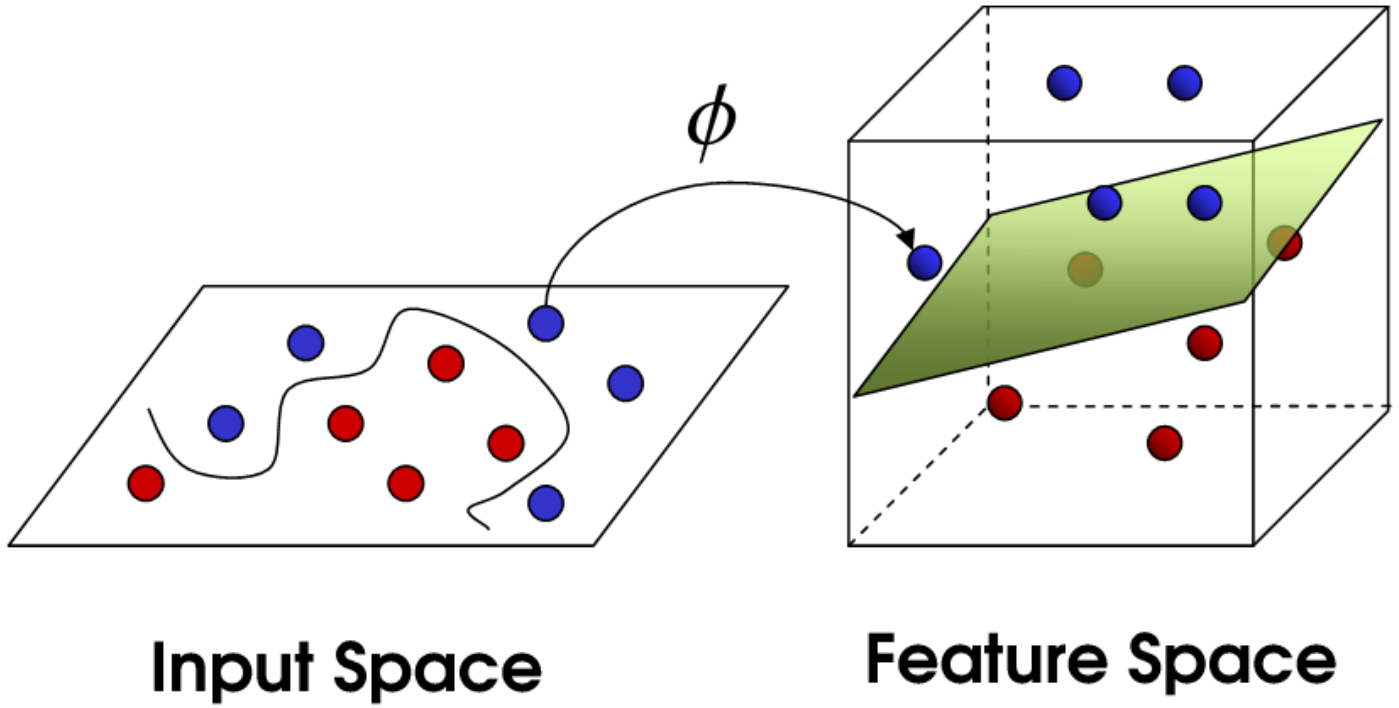
Snakes on a Hyperplane:
Python Machine Learning in Production



Jessica Lundin
Machine Learning Manager
Microsoft Research
@_JessicaLundin

https://notebooks.azure.com/LundinMachine

http://stackoverflow.com/questions/9480605/what-is-the-relation-between-the-number-of-support-vectors-and-training-data-and

What is machine learning?

“machine learning explores the study and construction of algorithms that can learn from and make predictions on data”

*Ron Kohavi; Foster Provost (1998). "Glossary of terms". Machine Learning. **30**: 271–274.*

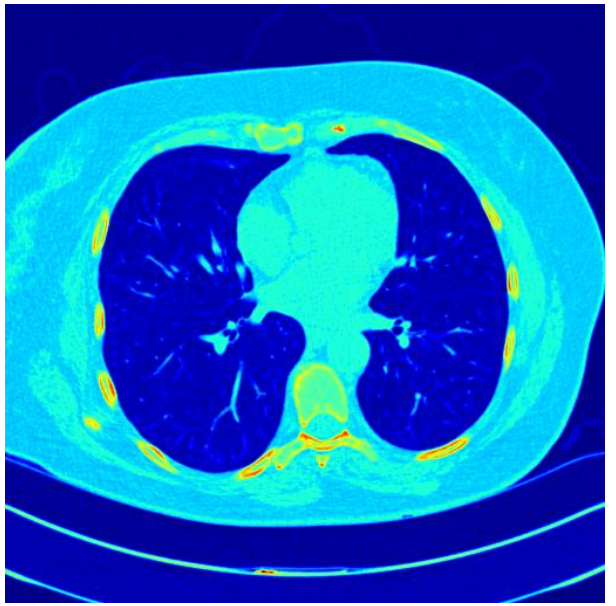
What is machine learning?

"machine learning explores the study and construction of algorithms that can learn from and make predictions on data"

Cat video classification

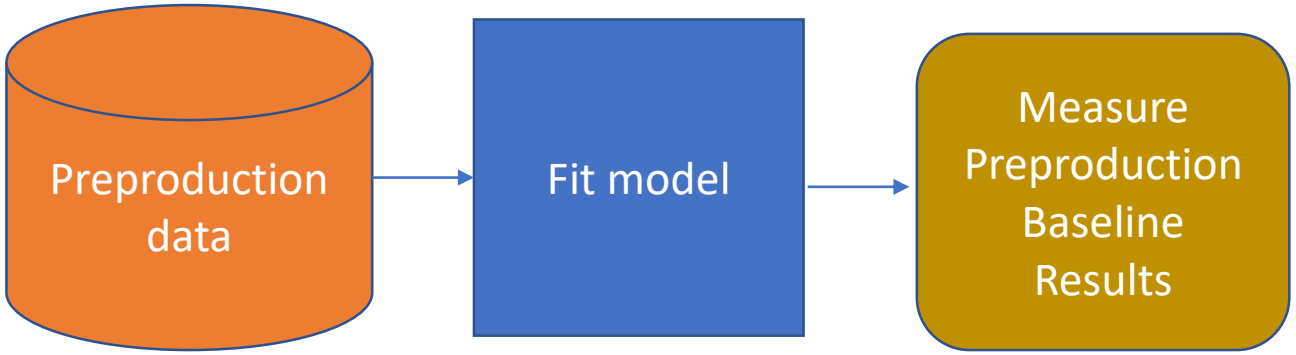
Handwritten digit identification

Lung cancer detection

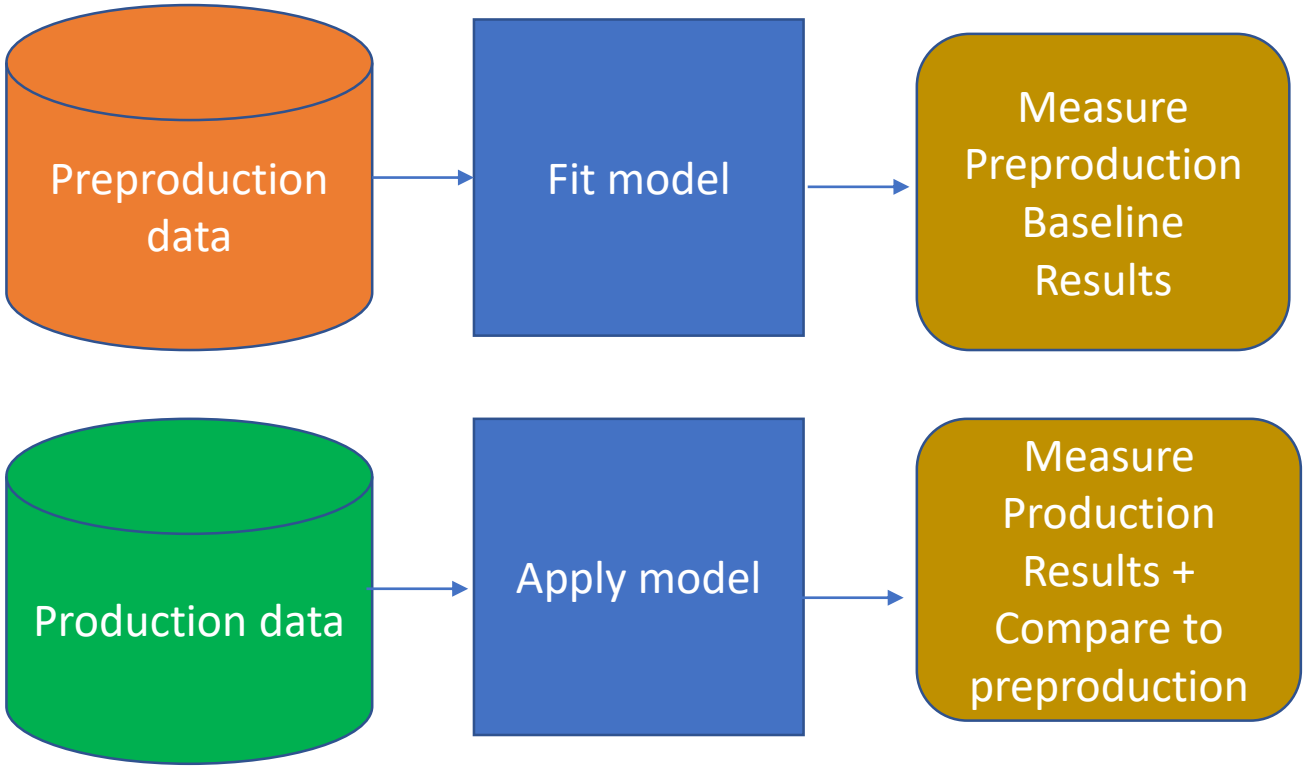


https://www.kaggle.com/c/data-science-bowl-2017/data

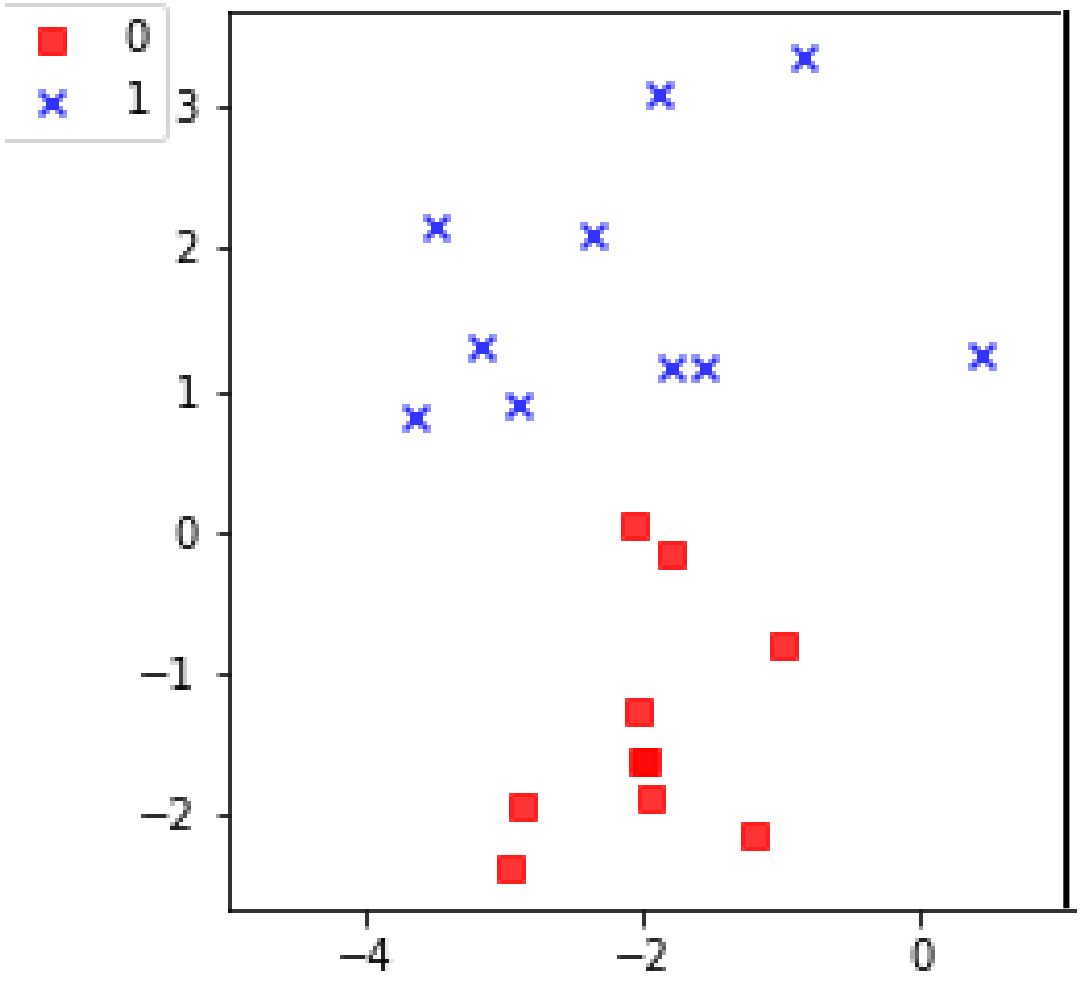
Machine learning in production: practical tips



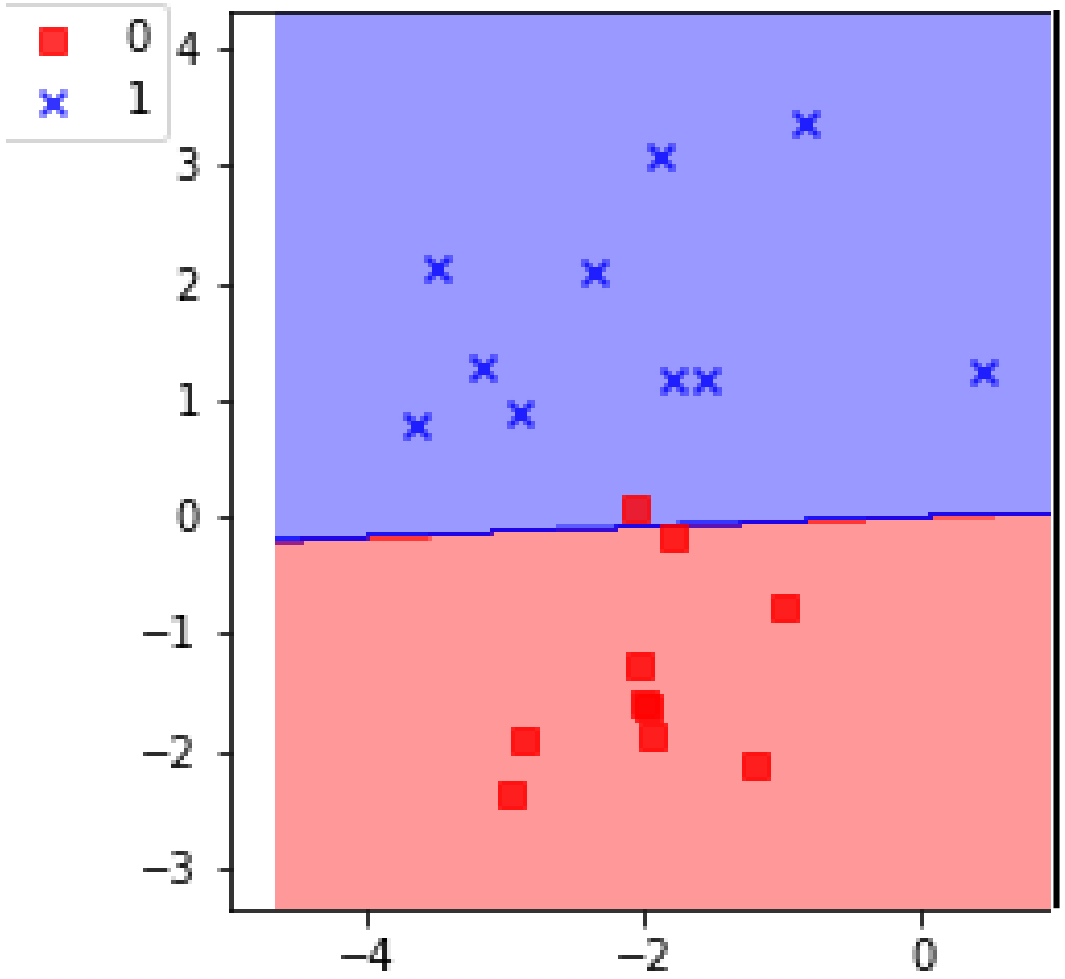
Machine learning in production: practical tips



Synthetic data



Synthetic data **fit a binary classifier**

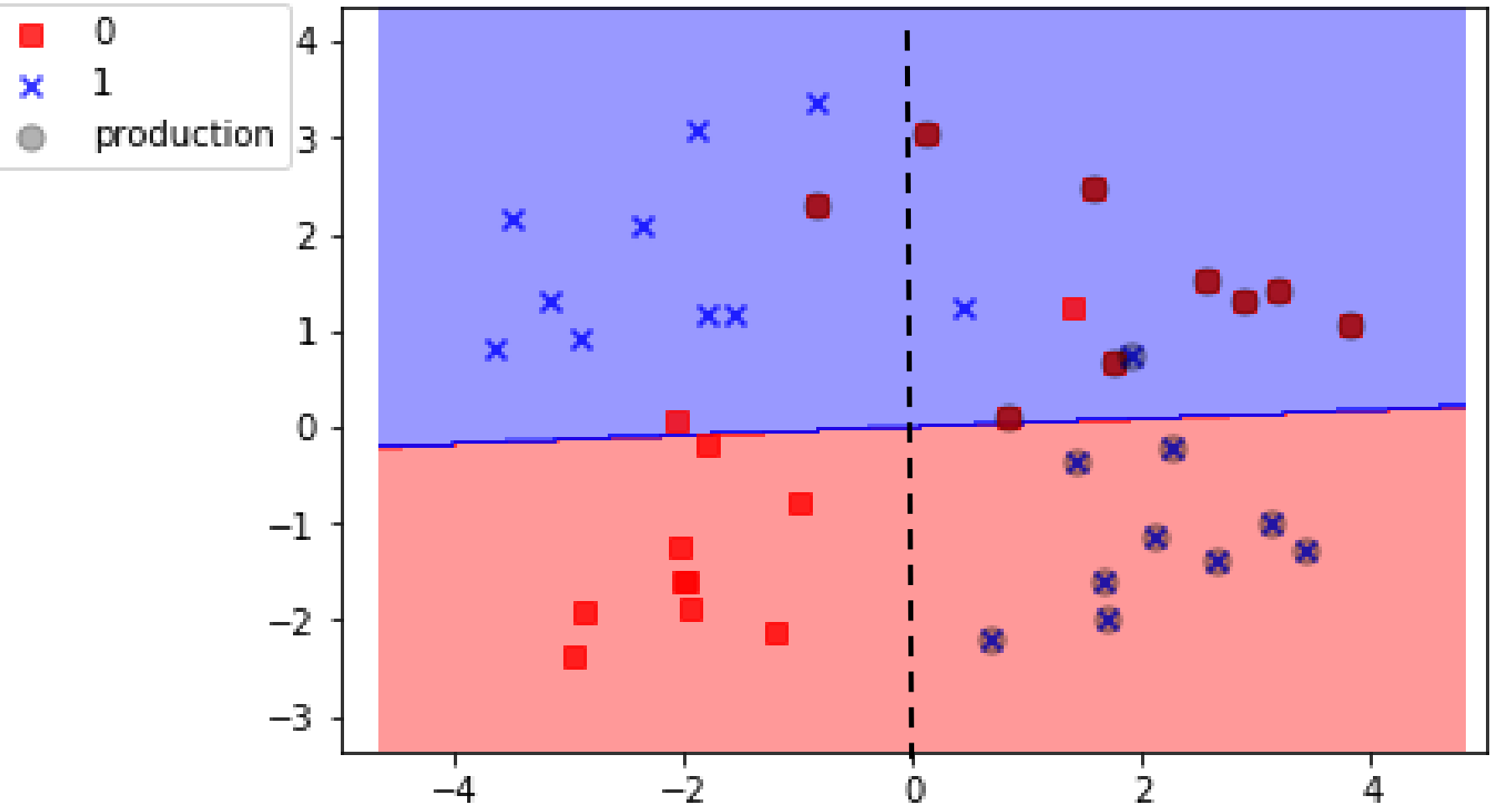


Preproduction
Accuracy = 95%

```
## Fit a Logistic Regression model
from sklearn.linear_model import LogisticRegressionCV
clf = LogisticRegressionCV()
clf.fit(X,y)
```

```
## measure the accuracy
clf.score(X,y)
```

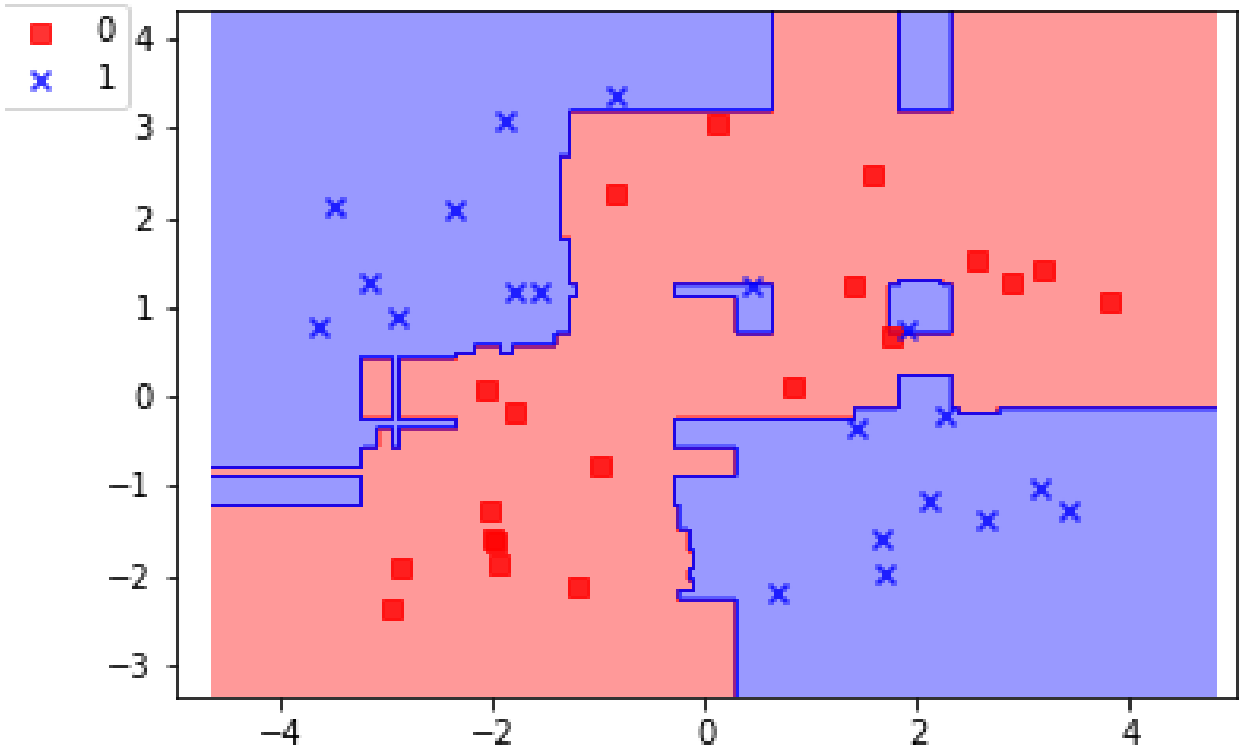
Unknown production distribution



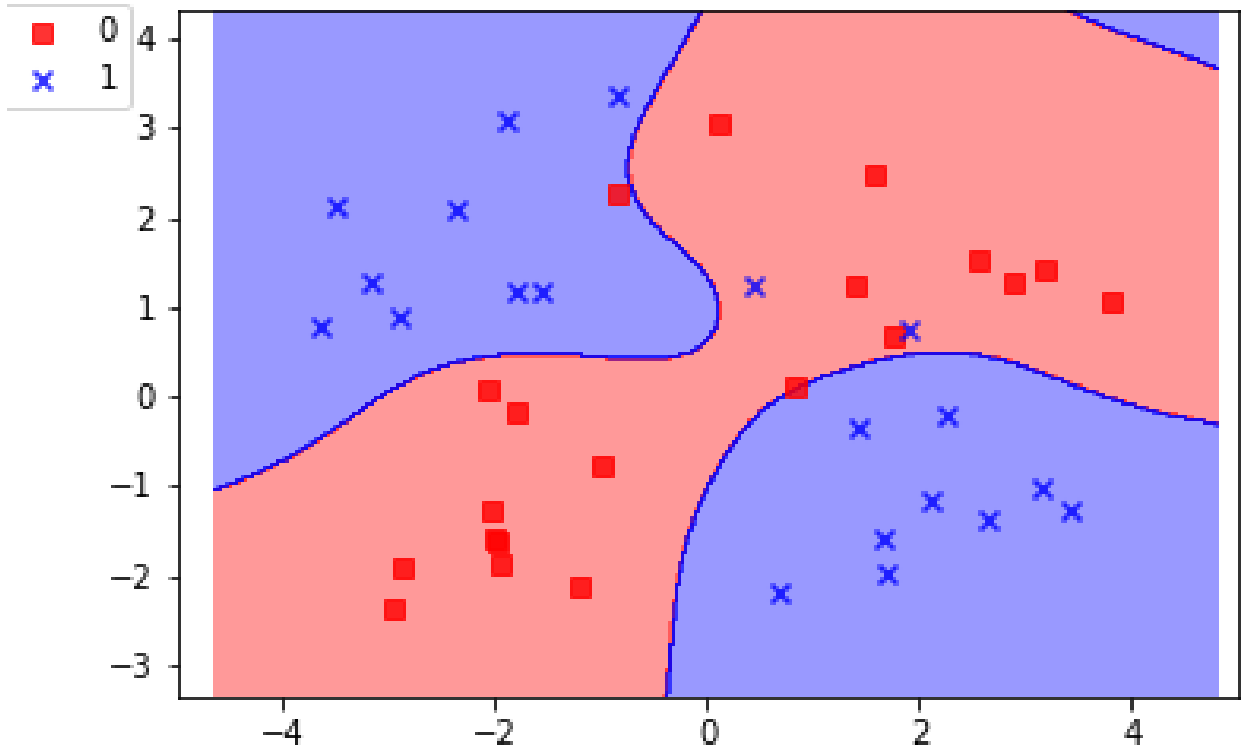
Preproduction
Accuracy = 95%

Production
Accuracy = 5%

Unknown production distribution
Retrain with non-linear algorithms



Random Forest
Accuracy = 100%

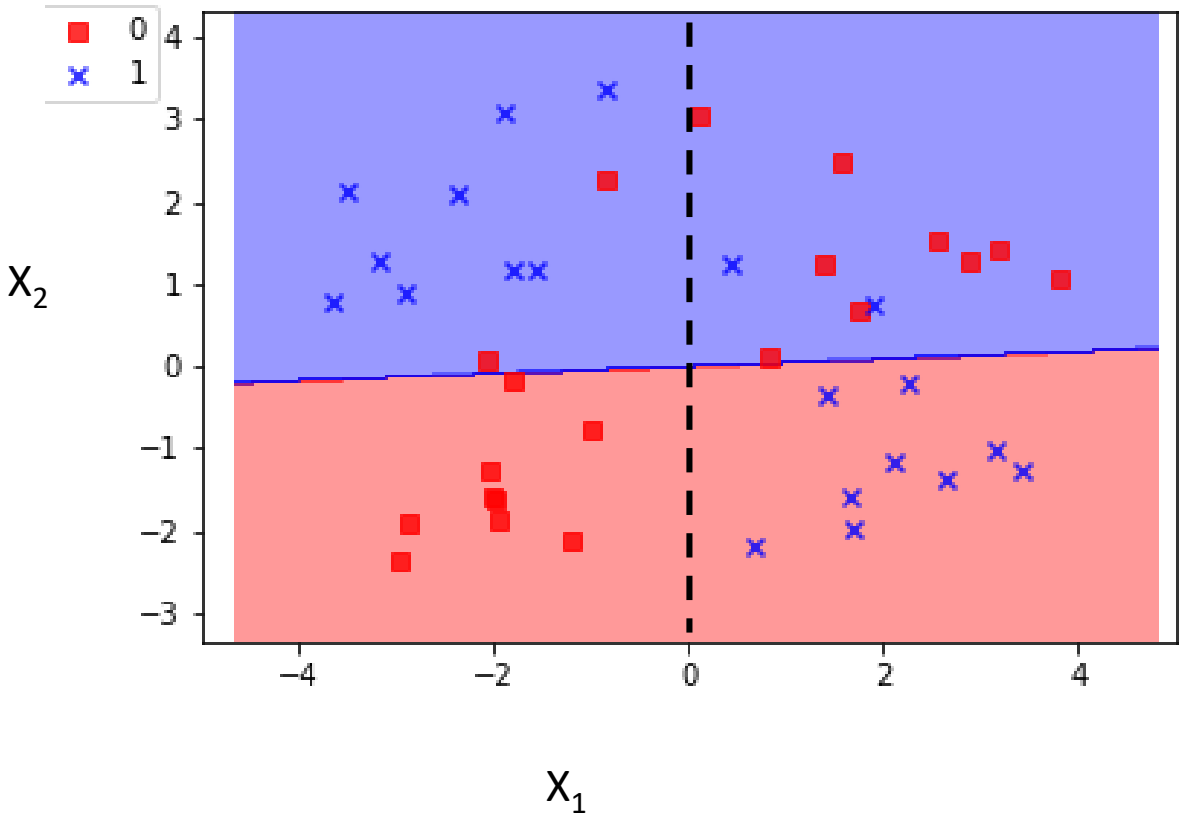


Support Vector Machine (SVM)
Kernel = Radial Basis Function
Accuracy = 93%

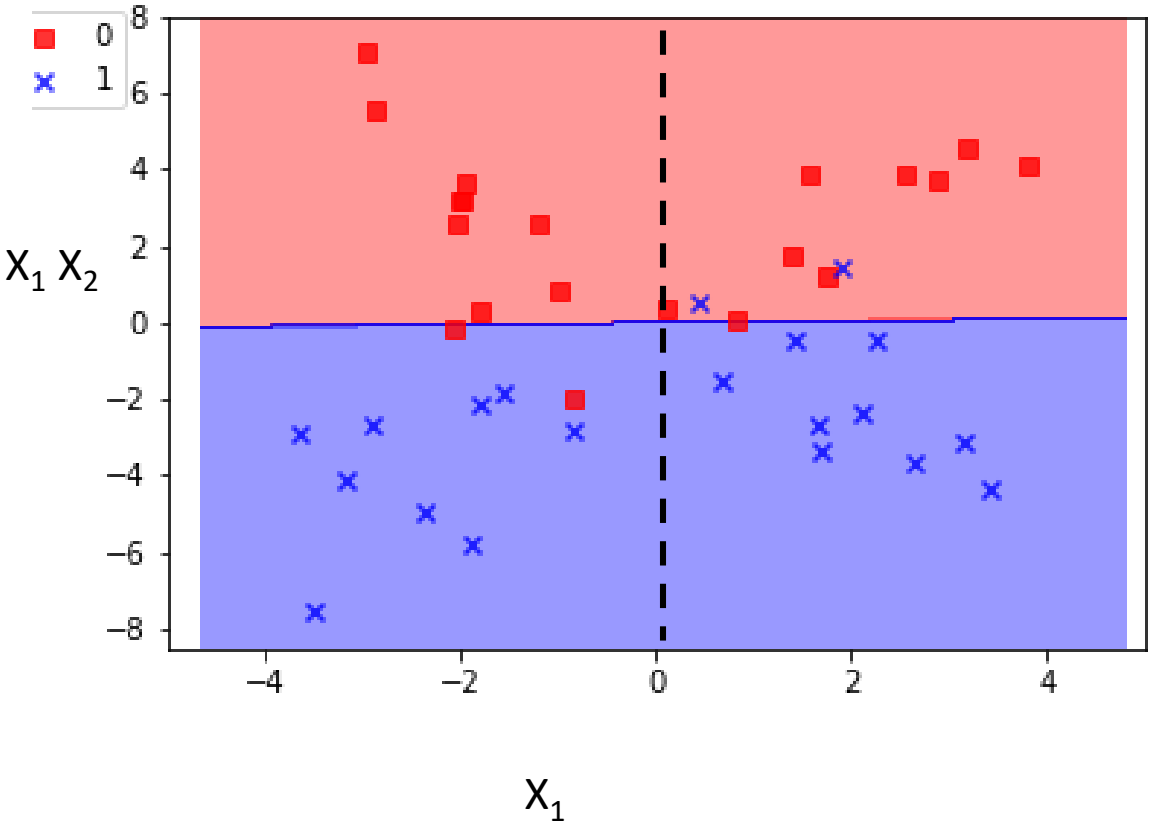
unknown production distribution

Feature engineering to linearize features

Original features



Modified features



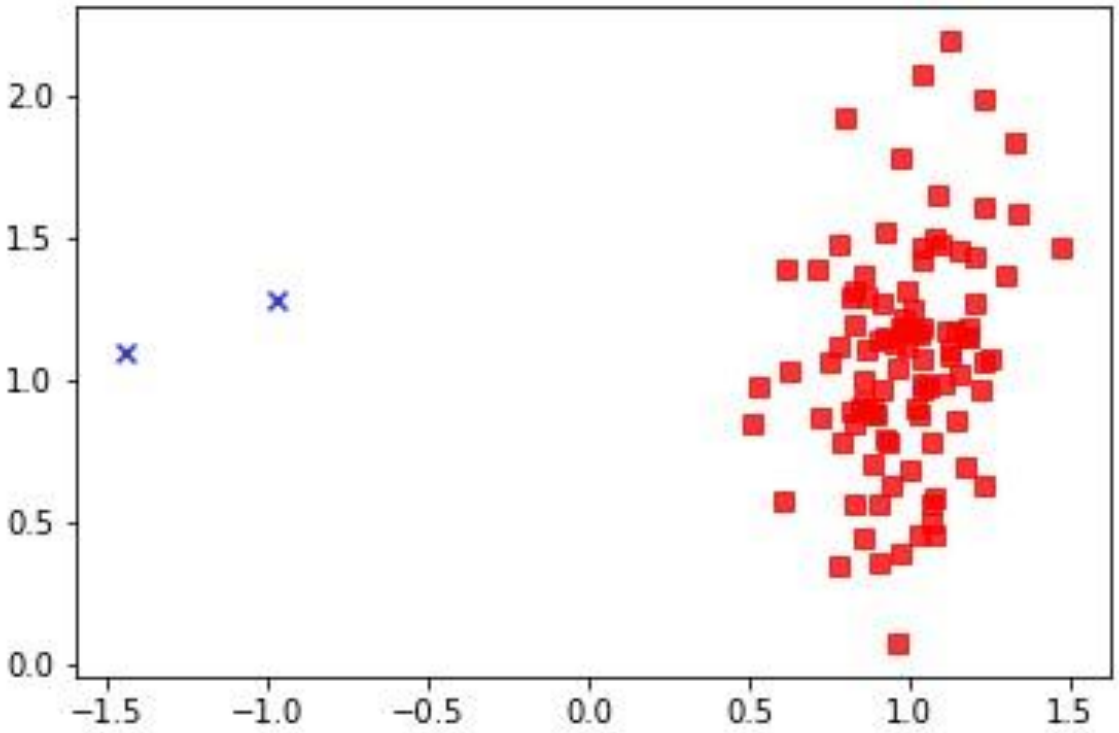
Accuracy = 90%

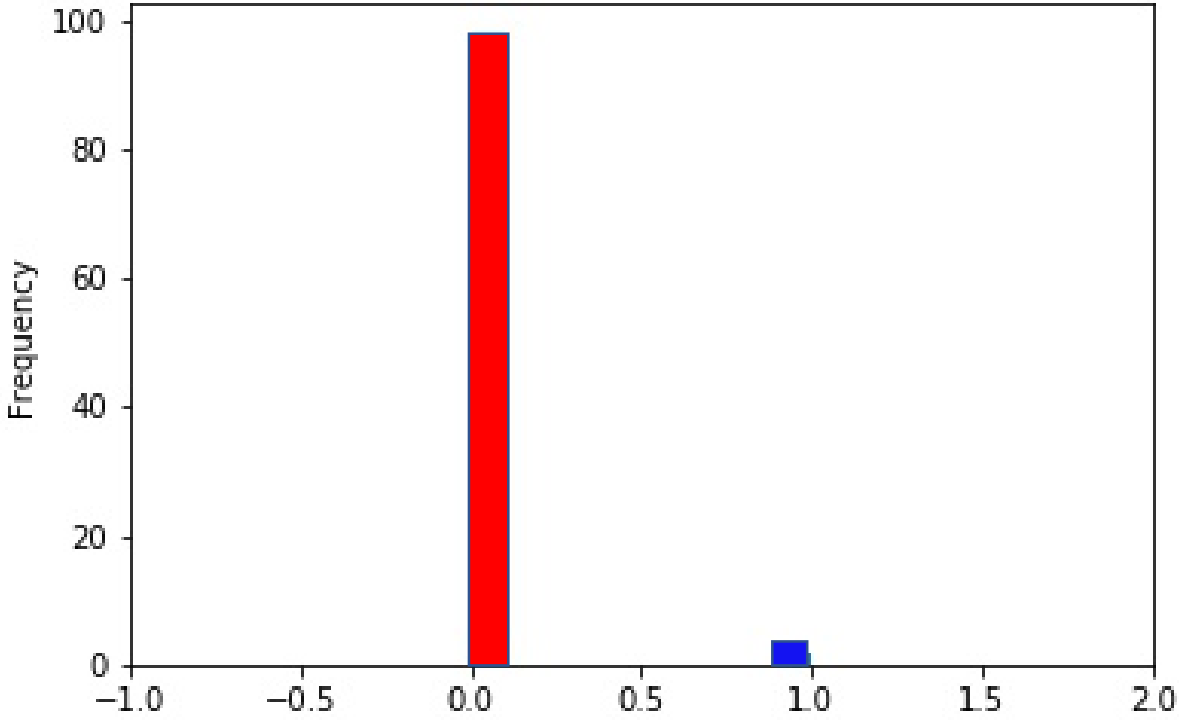
Model performance: unknown production distribution

Techniques for suspected distribution differences between preproduction and production:

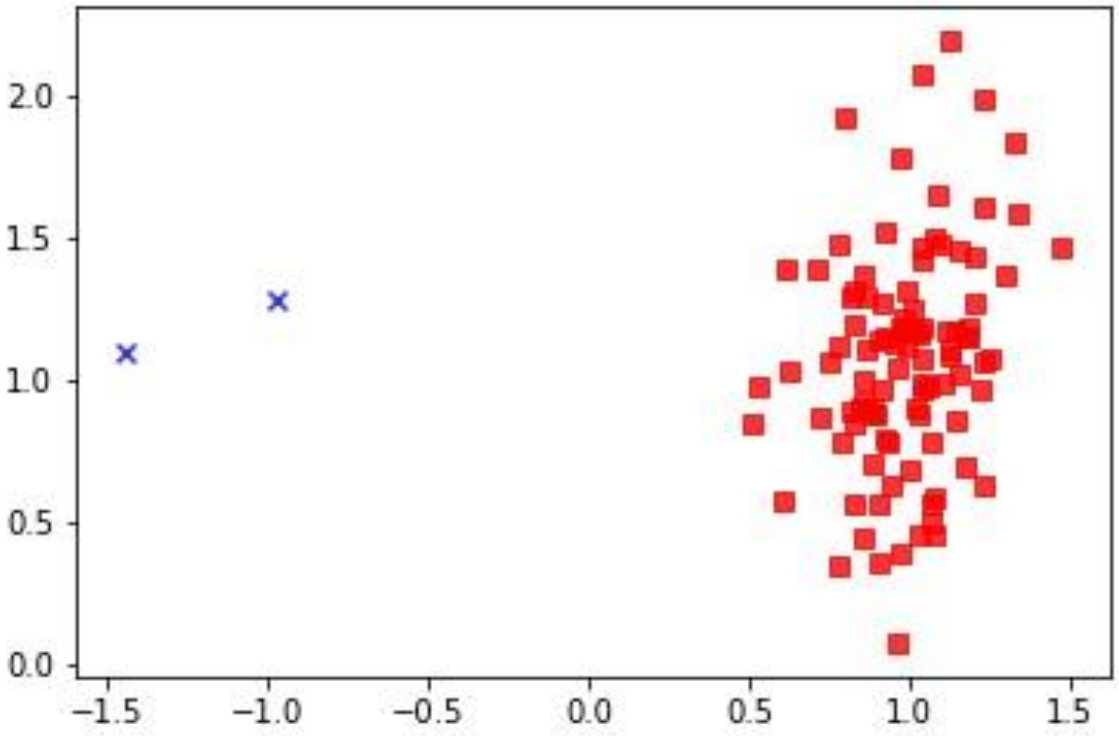
- Visualization (histograms, pairplots)
- Clustering
- Kullback-Leibler (KL) divergence

Model performance: unbalanced problems

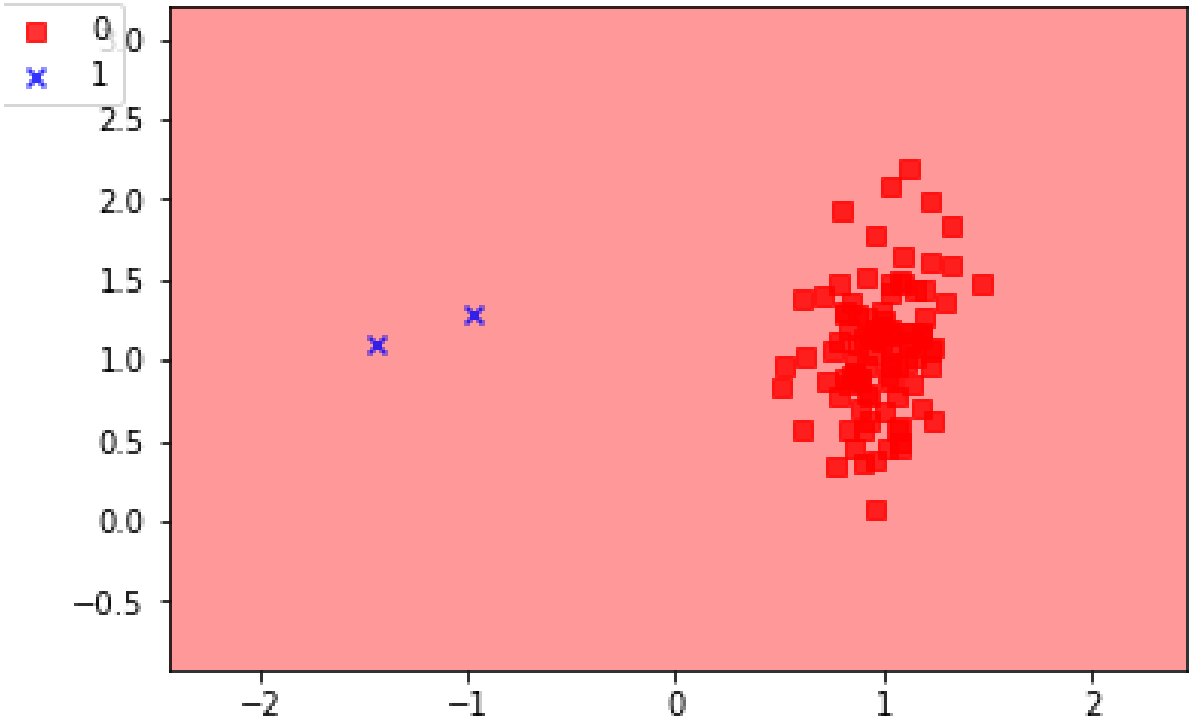




Model performance: unbalanced problems

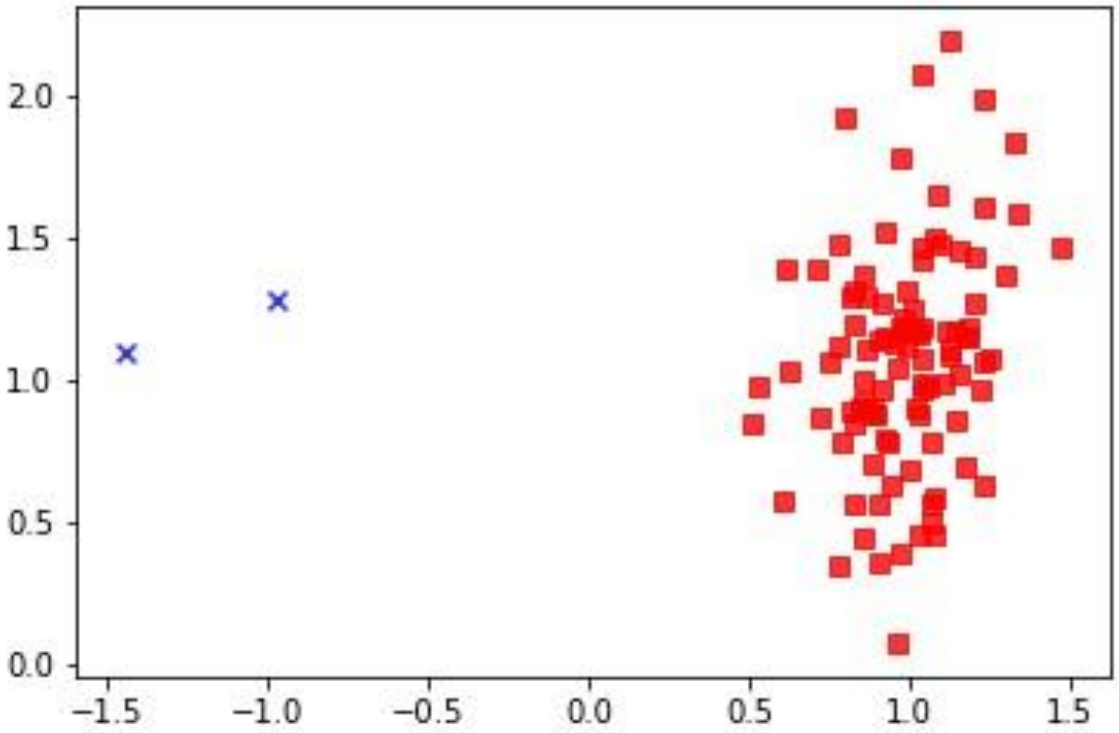


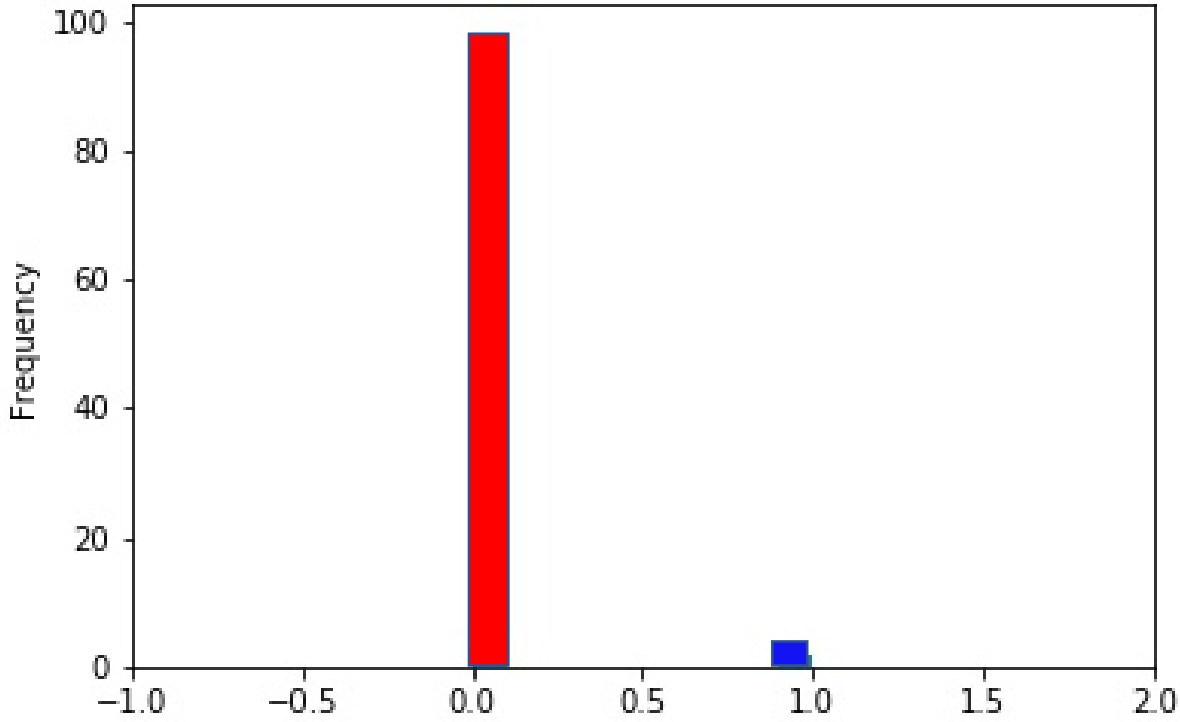
Model predicts single class 0 for all observations



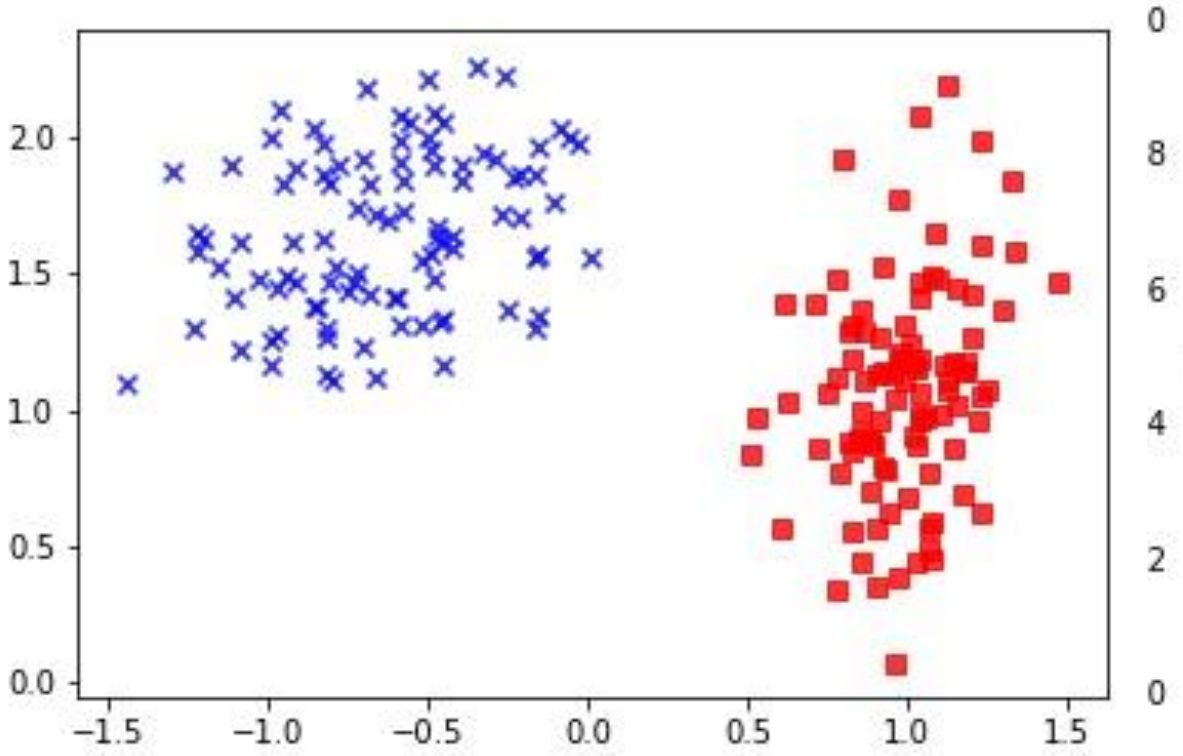
Accuracy: 0.98 = (Σ TP + Σ TN)/Σ total population
Precision: 0.0 = Σ TP/Σ prediction positive
Recall:0.0 = Σ TP/Σ condition positive

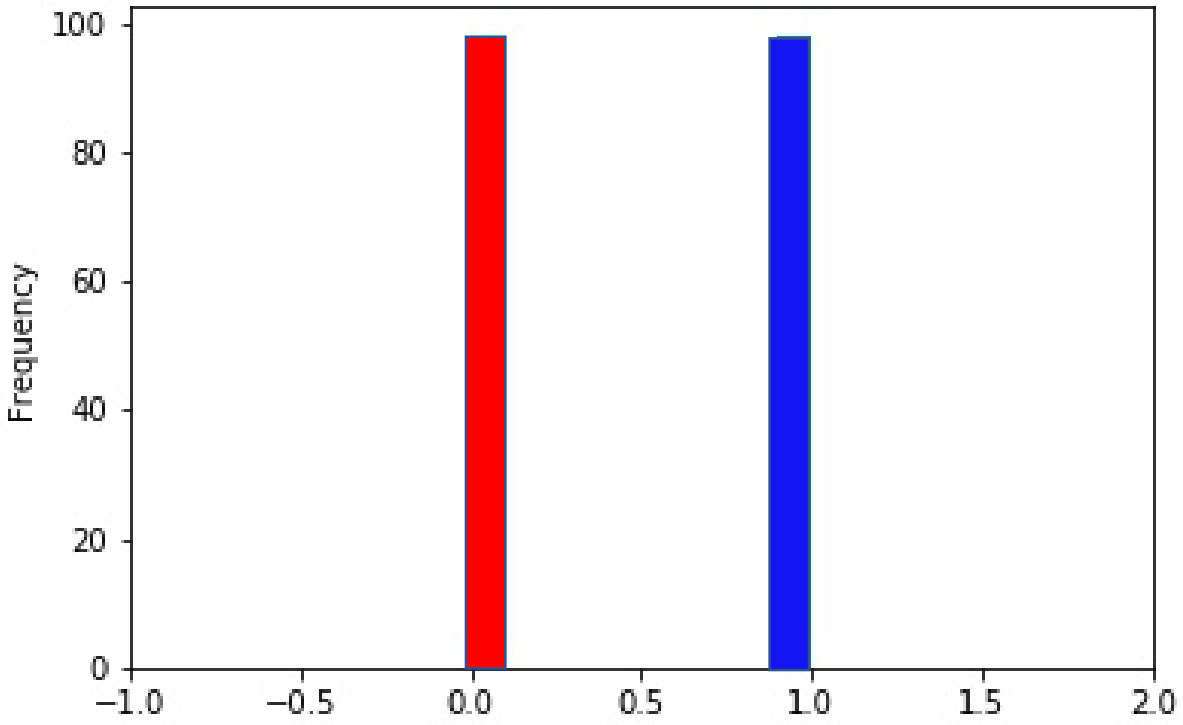
Model performance: unbalanced problems



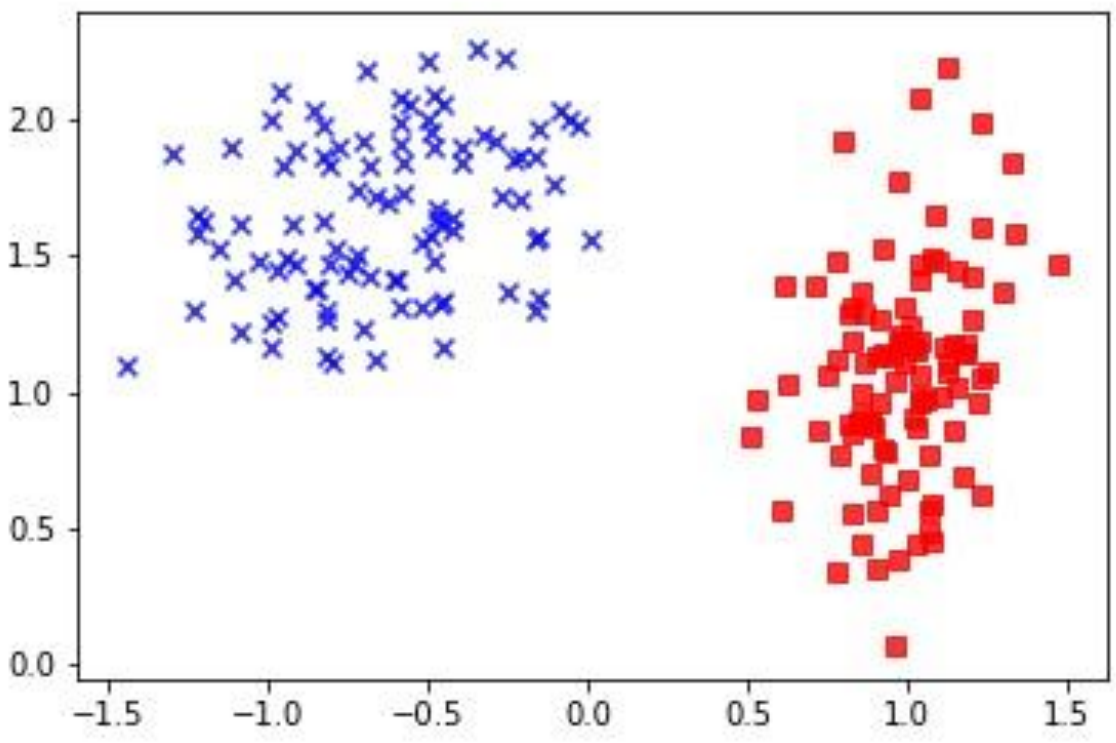


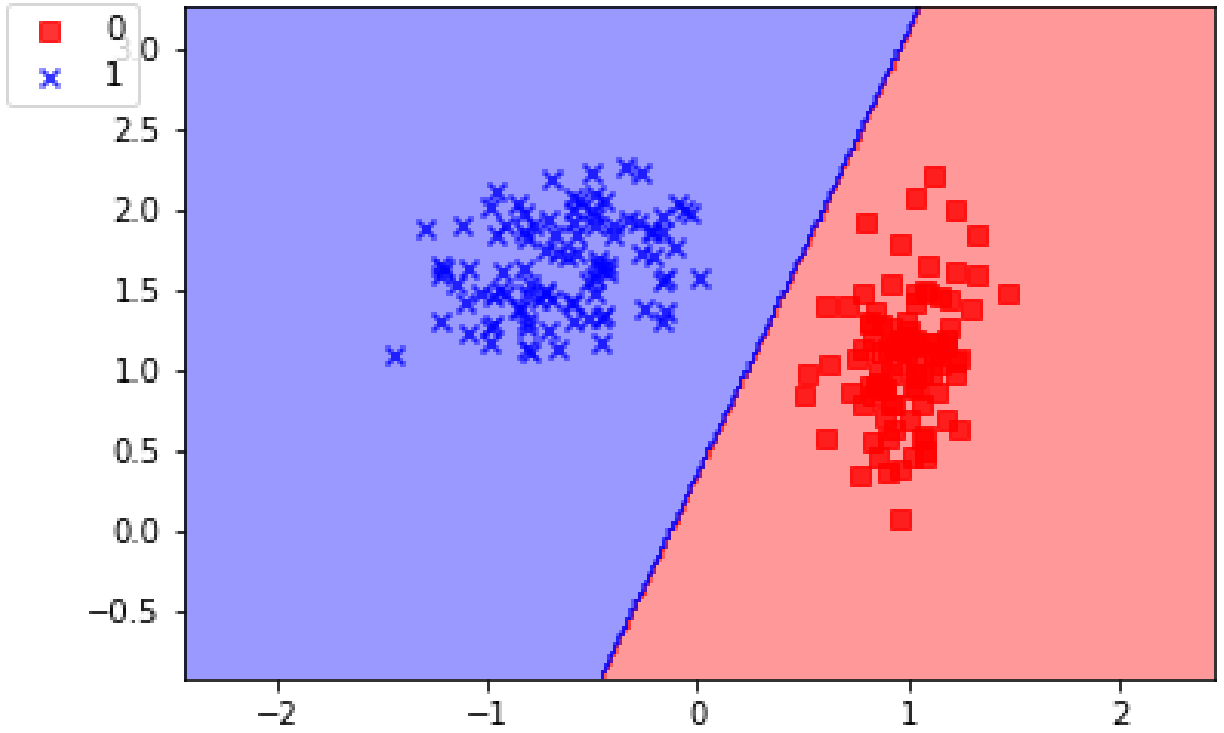
Model performance: unbalanced problems





Model performance: unbalanced problems





Accuracy: 1

Precision: 1 = Σ TP/Σ prediction positive

Recall: 1 = Σ TP/Σ condition positive

Model performance: unbalanced problems

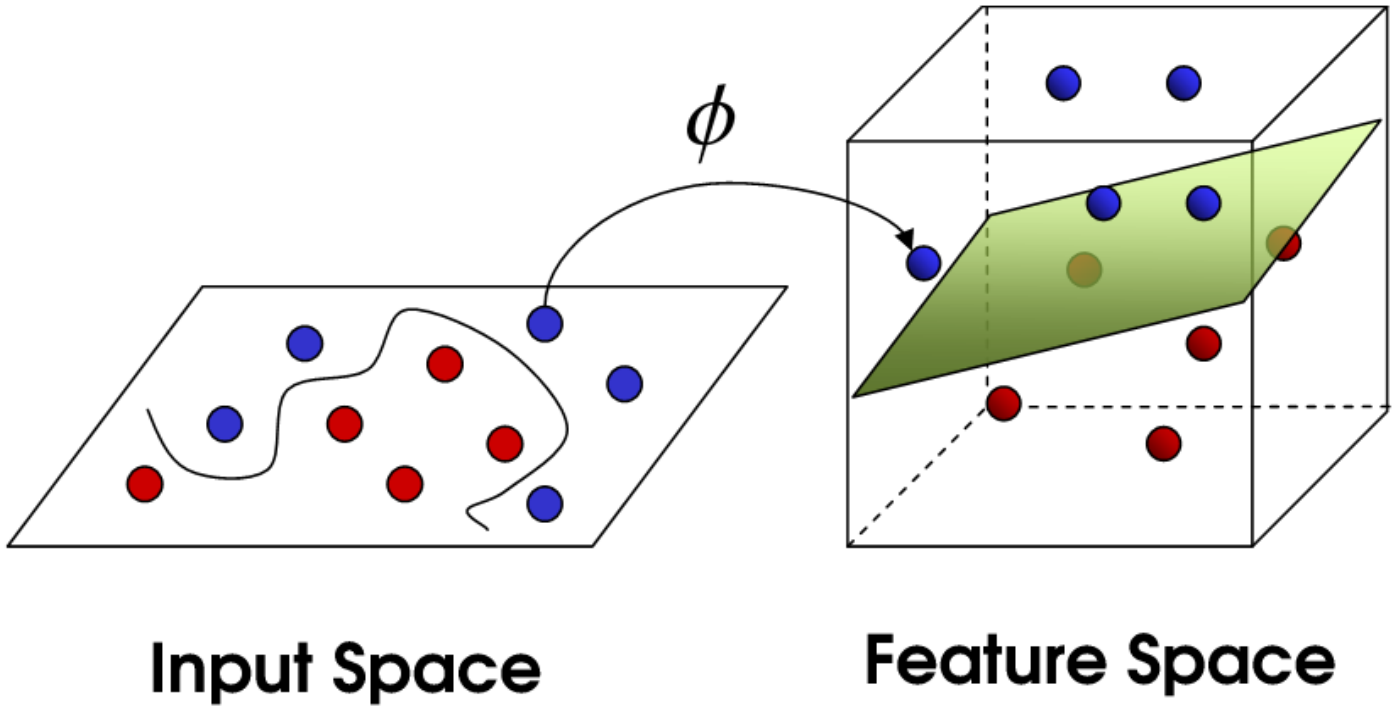
Techniques for unbalanced problems

Cost-sensitive classification:

- Rare-class upsampling with replacement
- Importance weighting
- Boosting

Treat it as an anomaly detection problem (one-class SVM)

Snakes on a Hyperplane:



http://stackoverflow.com/questions/9480605/what-is-the-relation-between-the-number-of-support-vectors-and-training-data-and

Machine learning in production: practical tips

Logging:

- Timestamp, Instance ids
- Model run time
- Model results, performance metrics
- Model convergence errors

Auditing:

- Manual process of digging into logs and data to resolve unexpected behavior

Machine Learning Resources

General Resources:

Introduction to Machine Learning, Coursera
by Andrew Ng
https://www.coursera.org/learn/machine-learning

The Elements of Statistical Learning
(free pdf download)
by Hastie, Tibshirani, Friedman
http://statweb.stanford.edu/~tibs/ElemStatLearn/

Kaggle Tutorials
https://www.kaggle.com/wiki/Tutorials

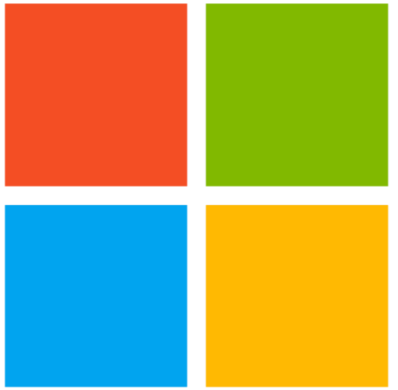
ML in Python:

Scikit Learn
http://scikit-learn.org/

Caffe
TensorFlow
CNTK
Theano
Keras
(packages all on github)

Rpy2: Python's R wrapper

Microsoft Python resources



Azure SDK - https://azure.microsoft.com/en-us/develop/python/

Intro to Python Programming - https://mva.microsoft.com/en-us/training-courses/introduction-to-programming-with-python-8360

Python tools for Visual Studio - https://microsoft.github.io/PTVS/

Cognitive Toolkit (CNTK) - https://www.microsoft.com/en-us/research/product/cognitive-toolkit/

Thanks!

Health ML team is hiring Data Scientists!
Come work at Microsoft Research

https://careers.microsoft.com/
ML/Data Scientist: 1030519
Developer: 1048462, 1032009, 1031571, 1031704, 1026221

@_JessicaLundin