Impact of Outliers On Supervised Machine Learning Classifiers

...

Online Encyclopedia of Integer Sequences

Github repository link: https://github.com/HeTalksInMaths/SupMLProj-OEIS

Project Overview

Given an integer sequence, use first nine values as features to determine if tenth value is larger than ninth.

Classifier use case - determine if multiple models required for some downstream task.

Which algorithms suffer when there are outliers?

Which algorithms generalize well to outliers?

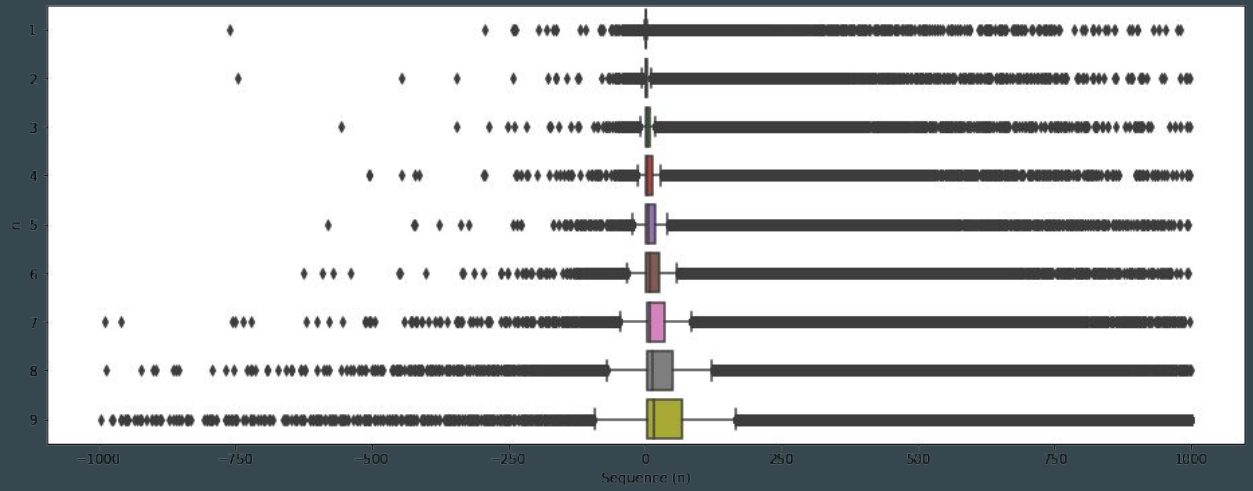
Imbalanced data sets - consider various metrics beyond accuracy.

Data

Online Encyclopedia of Integer Sequences (OEIS), 350,000 + seqs.

Large data set first nine values between -10^{10} and 10^{10}, < 300,000 seqs.

Small data set first nine values between -1000 and 1000, < 200,000 seqs.



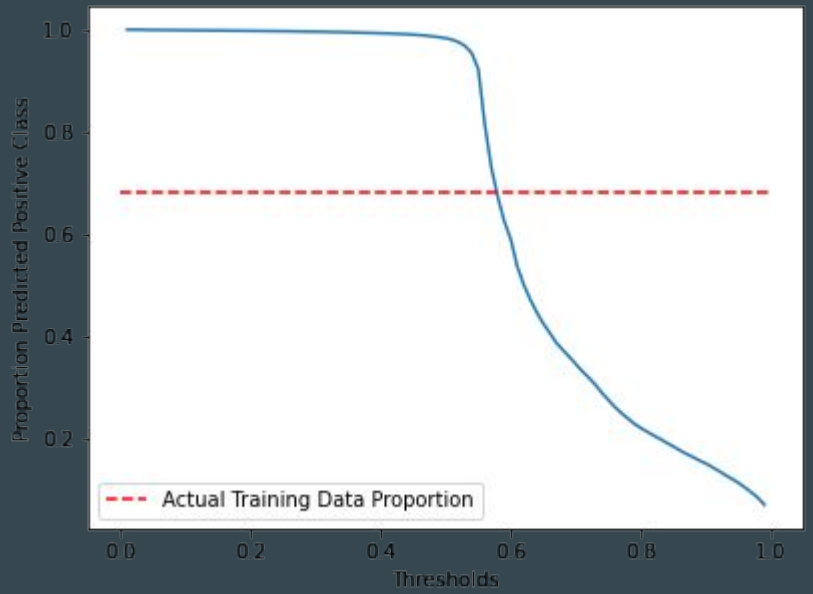
Models

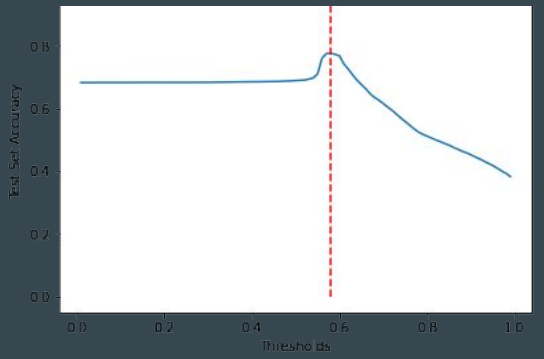
Logistic regression with threshold moving hyperparameter tuning.

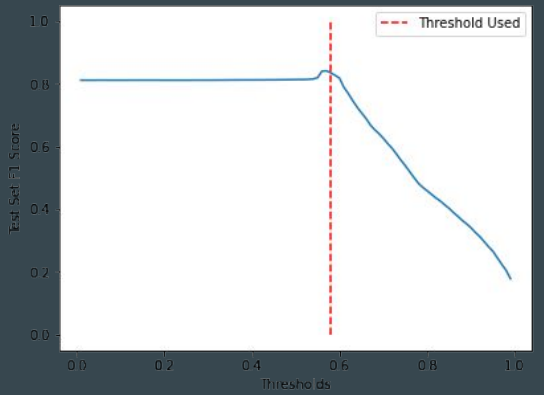
AdaBoost, 100 estimators.

XGBoost, 100 estimators.

Threshold Moving (Logistic Regression)







Results

Large data set

Metric	All Pos. Bench.	LR w/ outliers	LR w/ thresh. adj.	AdaBoost, n_est=100	AdaBoost small to large	XGBoost, n_est=100	XGBoost small to large
Accuracy	0.760	0.759	0.825	0.839	0.708	0.860	0.863
F_1	0.865	0.855	0.885	0.896	0.793	0.910	0.911
AUC	0.500	0.787	0.834	0.865	0.747	0.898	0.886

Small data set

Metric	All Pos. Bench.	LR w/ thresh. adj.	AdaBoost, n_est=100	XGBoost, n_est=100
Accuracy	0.682	0.776	0.790	0.826
F_1	0.811	0.835	0.850	0.874
AUC	0.500	0.821	0.845	0.890

Key Takeaways (for OEIS)

Logistic regression performance susceptible to outliers but reduced data model can still generalize to outliers.

Threshold moving can salvage a model if use case dependent on predicting hard class labels.

AdaBoost robust to outliers but cannot generalize from reduced data model to outliers.

XGBoost performs best, robust to outliers and achieves out-of-distribution generalizability.