SHARCNET™

Train machine learning models to predict rare events

Weiguang Guan (guanw@sharcnet.ca)

SHARCNet/Alliance

Compute Ontario

Examples of rare event prediction

Binary classifications with skewed ratio:

- Fraud detection in financial transactions
- Cancer screening
- Extreme weather conditions (hale, hurricane, etc) forecast

Ratio of positives to negatives < 1:100

Training on imbalanced dataset

- Bias adjustment to reflect the skewed data distribution
- Weight adjustment of samples
- Sampling the original dataset
  - Either down-sampling negative samples
  - Or over-sampling positive samples
- Combine weight adjustment and sampling strategies

Output of prediction model

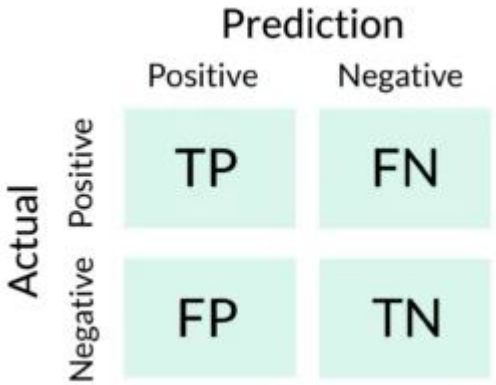
Output of model

A real-valued number representing score or probability of being positive

Final classification

Thresholding the output: 1 if output>T, or 0 otherwise.

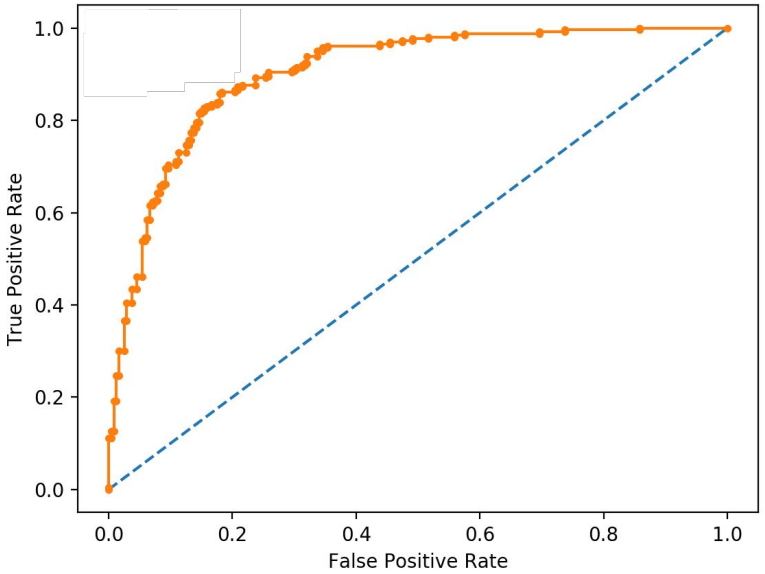
Performance measurement



Metrics

- Accuracy = (TP+TN) / (P+N)
- True positive rate (TPR) = TP/P = TP / (TP+FN)
- False positive rate (FPR) = FP/N = FP / (TN+FP)
- Recall is same as TPR
- Precision = TP / (TP+FP)
- Area under the curve (AUC) is the area under the ROC defined by TRP vs FRP

Operating point on ROC



Some applications (cancer screening, etc)

- Prefer higher TPR
- Tolerate higher FPR

Others (fingerprint recognition as authentication)

- Prefer lower FPR
- Tolerate lower TPR

Jupyter notebook

Data: Kaggle competition “Credit Card Fraud Detection” at
https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Notebook: Tensorflow tutorial “Classification on imbalanced data” at
https://www.tensorflow.org/tutorials/structured_data/imbalanced_data