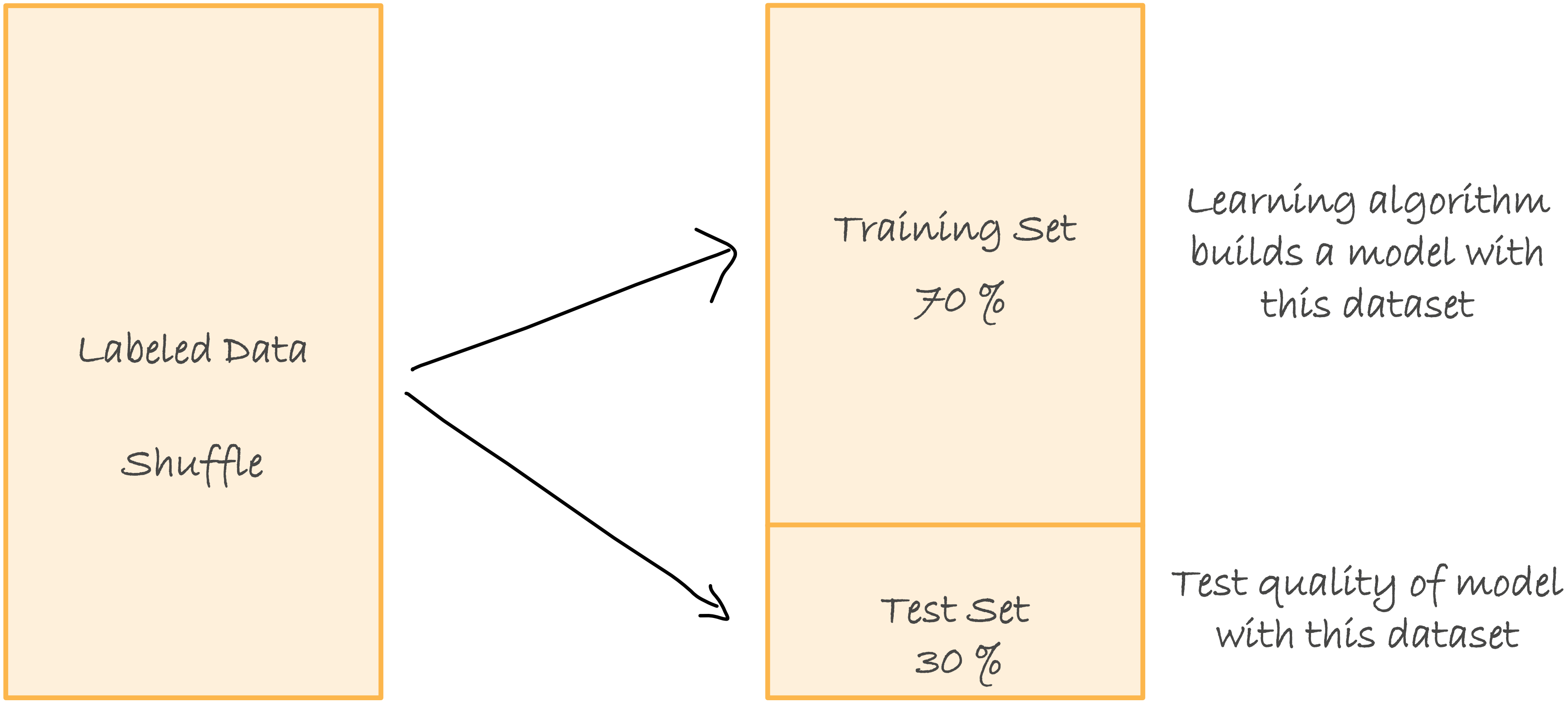
Model Performance

Quality Metrics

Content Prepared By: Chandra Lingam, Cloud Wave LLC

Terminologies



Verify Model Fit

Fit	Issue
Underfitting	Poor performance on both training and test set
Overfitting	Good performance on training set Poor performance on test set
Balanced	Good performance on training set and test set

Supervised Algorithm Types

Algorithm	Output
Regression	Continuous Numeric
Binary	Binary
Multi-class Classification	Categorical – One of many possible known outcomes

Regression

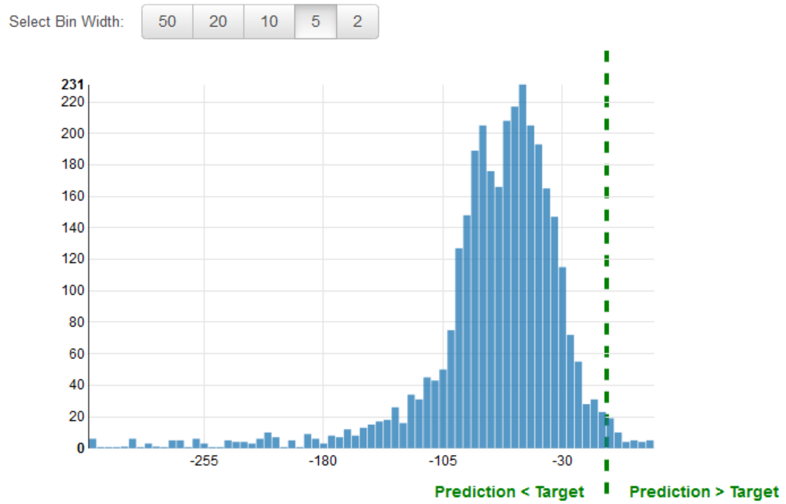
Root Mean Square Error (RMSE)

$$RMSE = \sqrt{1/N \sum_{i=1}^{N} (actual\ target - predicted\ target)^2}$$

https://docs.aws.amazon.com/machine-learning/latest/dg/regression-model-insights.html

Regression

Residual Histogram



https://docs.aws.amazon.com/machine-learning/latest/dg/regression-model-insights.html

Difference = Predicted – Target

Plot the difference as a histogram

A good model has balanced over and under predictions

Binary Classifier

Positive Class – Condition that we are interested in detecting

Admit Student?

Positive Class = *Admitted*

Negative Class = *Not-Admitted*

Binary Classifier

Positive Class – Condition that we are interested in detecting

Is this patient at risk of developing heart disease?

Positive Class = Heart Disease

Negative Class = Normal

Binary Classifier

Positive Class – Condition that we are interested in detecting

Is this email a SPAM?

Positive Class = Spam

Negative Class = Normal

Binary Classifier Output

Some Algorithms directly produce a binary output

Some Algorithms produce a raw-score that gives a probability of an example belonging to positive class

Raw Score to Binary Class

Raw Score	Class
0	Negative
0.1	Negative
0.2	Negative
0.3	Negative
0.4	Negative
0.5	Positive
0.6	Positive
0.7	Positive
0.8	Positive
0.9	Positive
1	Positive

Cut-off, Threshold

Email Spam Classifier

Raw Score	Class
0	Negative
0.1	Negative
0.2	Negative
0.3	Negative
0.4	Negative
0.5	Negative
0.6	Negative
0.7	Negative
0.8	Positive
0.9	Positive
1	Positive

Positive: Spam
Negative: Normal

Increasing Cut-off

1. Reduces possibility of a Normal email marked as spam (Reduces False Positive)
2. Increases possibility of a Spam marked as normal email (Increases False Negative)

Cut-off, Threshold

Identify Patients at risk for a disease

Raw Score	Class
0	Negative
0.1	Negative
0.2	Negative
0.3	Positive
0.4	Positive
0.5	Positive
0.6	Positive
0.7	Positive
0.8	Positive
0.9	Positive
1	Positive

Cut-off, Threshold

Positive: At-Risk
Negative: Normal

Lowering Cut-off

1. Reduces possibility of missing an at-risk patient (Increased Recall)
2. Increases possibility of normal person flagged as at-risk (Increase in False-Alarm)

Confusion Matrix - Concepts

Terminology	Description
Positive	Total Actual Positives = True Positive + False Negative
Negative	Total Actual Negatives = True Negative + False Positive
True Positive	How many samples were correctly classified as Positive
True Negative	How many samples were correctly classified as Negative
False Negative	How many positive samples were mis-classified as negative
False Positive	How many negative samples were mis-classified as positive

True Positive Rate

Fraction of Positives predicted correctly

$$TPR = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

TPR is also referred as Recall, Probability of detection

Model with Recall closer to 1 is good. Model with Recall closer to 0 is poor.

True Negative Rate

Fraction of Negatives predicted correctly

$$TNR = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

Model with TNR closer to 1 is good. Model with TNR closer to 0 is poor.

False Positive Rate

Fraction of negatives mis-classified as positives.

$$FPR = \frac{False\ Positive}{True\ Negative + False\ Positive}$$

FPR is also referred as Probability of false alarm

Model with FPR closer to 0 is good. Model with FPR closer to 1 is poor.

False Negative Rate

Fraction of positives mis-classified as negatives.

$$FNR = \frac{False\ Negative}{True\ Positive + False\ Negative}$$

FNR is also referred as misses.

Model with FNR closer to 0 is good. Model with FNR closer to 1 is poor.

Precision

Fraction of true positives among all predicted positives. Larger value indicates better predictive accuracy

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Good model has precision closer to 1. Poor model has precision closer to 0

Accuracy

Fraction of correct predictions. Larger value indicates better predictive accuracy

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Negative + False\ Positive + True Negative}$$

Good model has accuracy closer to 1

F1 Score

Harmonic Mean of Precision and Recall

$$F1\ Score = \frac{2 * Precision * Recall}{Precsion + Recall}$$

Good model has F1 score closer to 1

Area Under Curve (AUC)

AUC is the area of a curve formed by plotting True Positive Rate against False Positive Rate at different cut-off thresholds

Good model have AUC closer to 1

0.5 is considered random guess

Closer to 0 is unusual and it indicates model is flipping results

Summary

Don't depend on one metric

For classification, recall (true positive rate) and precision along with F1 Score are often used together

Corner Cases:

Recall = very high when a model classifies all samples as positive

Precision = very high when a model classifies one positive correctly and misclassifies all other samples as negative

F1 Score = balances these corner cases