outrageously
AMBITIOUS

Module 3: Evaluating & Interpreting Models

Duke
PRATT SCHOOL *of* ENGINEERING

Module 3 Objectives:

At the conclusion of this module, you should be able to:

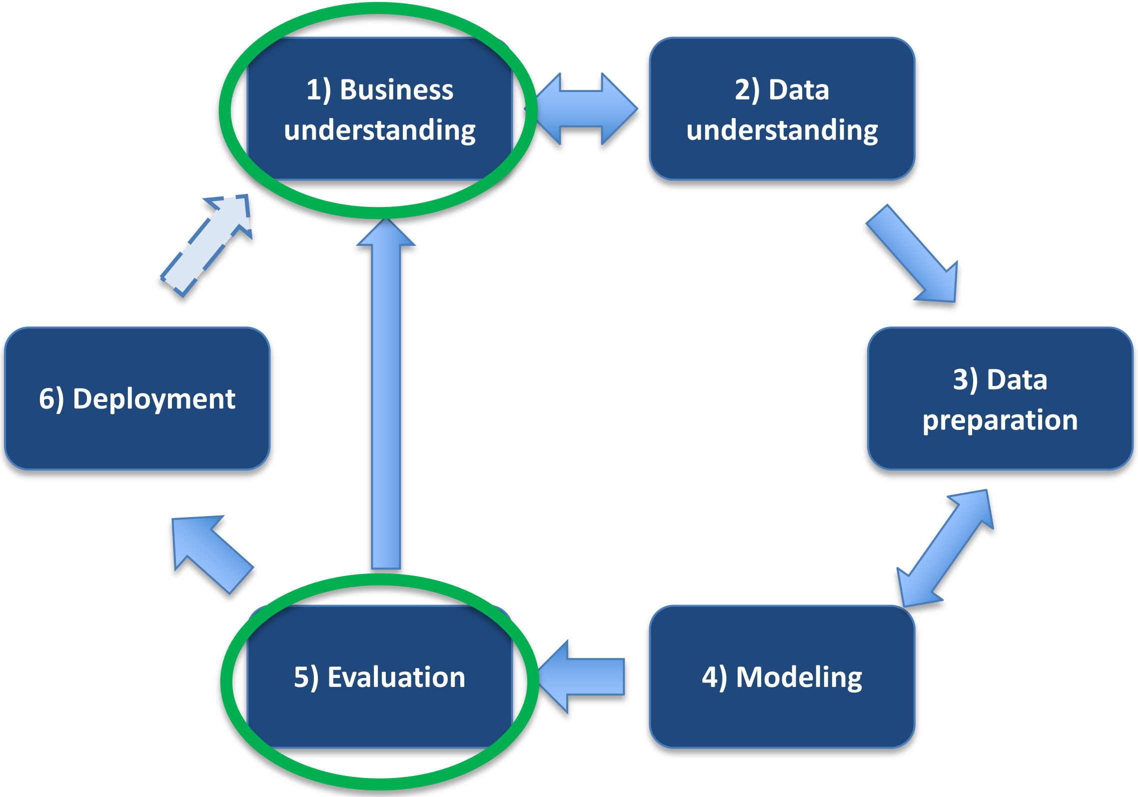
1) Differentiate between outcome and output metrics
2) Apply metrics to evaluate the performance of regression models
3) Apply metrics to evaluate the performance of classification models

outrageously
AMBITIOUS

Outcomes vs. Outputs

Duke
PRATT SCHOOL of
ENGINEERING

Evaluating Models



Outcomes vs. Outputs

Outcome

- Refers to the desired business impact on your organization or for your customer
- Stated in terms of the expected impact (which is often $)
- Does NOT contain model performance metrics or other technical metrics

Outcomes vs. Outputs

Output

- Refers to the desired output from the model
- Measured in terms of a model performance metric
- Typically not communicated to the customer
- Set this AFTER setting the desired outcome

Outcomes vs. Outputs

	A tool to predict turbulence for airlines	A power demand forecasting tool for a utility
Outcome	• Low # of safety incidents per year, or lower $ of safety-related claims	• Lower cost per MWh of power produced • Lower emissions rate per MWh
Output	• Classification error metric (binary or 1-5 scale)	• Regression error metric

outraguously AMBITIOUS

Model Output Metrics

Duke
PRATT SCHOOL *of* ENGINEERING

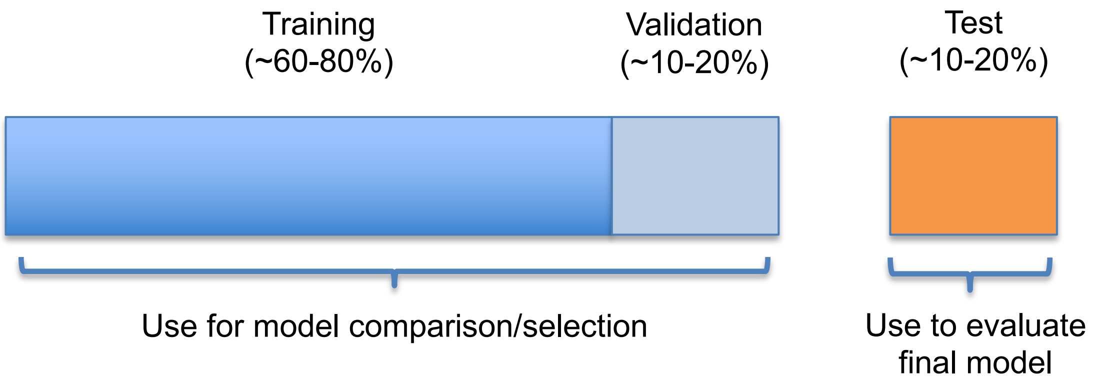
Uses for Metrics

We use metrics at several points in modeling:

- Model comparison & selection
- Evaluating model to deploy
- Ongoing model performance monitoring

Evaluating Models

- When comparing models, we calculate metrics using the **validation set** (or cross-validation)
- When evaluating our final model, we use the **test set**



Regression Error Metrics

MSE, MAE and MAPE

Mean Squared Error

$$MSE = \frac{1}{n}\sum_{i}(y_i - \hat{y}_i)^2$$

- Most popular regression error metric
- Heavily influenced by outliers - penalizes large errors heavily
- Influenced by scale of data
- Sometimes used as RMSE

MSE, MAE and MAPE

Mean Absolute Error

$$MAE = \frac{1}{n}\sum_{i} |y_i - \hat{y}_i|$$

- Also influenced by scale
- More robust to outliers
- Can be easier to interpret in context of the problem

MSE, MAE and MAPE

Mean Absolute % Error

$$MAPE = \frac{1}{n}\sum_{i}\frac{|y_i - \hat{y}_i|}{y_i}$$

- Converts error to a percentage
- Popular because it is easily understood
- Skewed by high % errors for low values of y

Example: MAE vs. MSE/RMSE

Case 1: Small variance in errors

Datapoint #	Error	|Error|	Error^2
1	1	1	1
2	1	1	1
3	1	1	1
4	2	2	4
5	2	2	4

Total Error: 7
MAE: 1.4 MSE: 2.2

Case 2: One large error

Datapoint #	Error	|Error|	Error^2
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	7	7	49

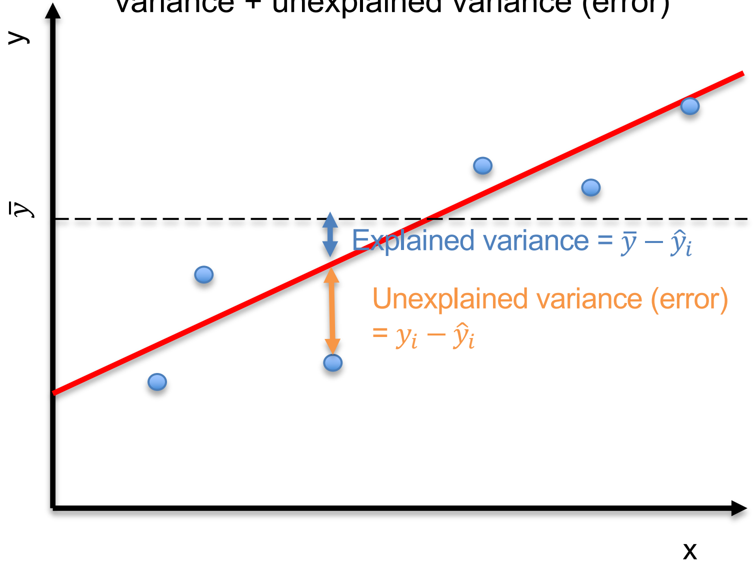
Total Error: 7
MAE: 1.4 MSE: 9.8

- MSE/RMSE penalizes severe errors much more than MAE
- This is sometimes desirable, as being off by a lot one time can be much worse than being off by a little every time

Coefficient of determination (R^2)

R-squared is used to communicate how much of the variability in your target variable (y) is explained by your model

Total deviation from mean = explained variance + unexplained variance (error)



Total squared deviation from mean (SST) = Sum of squares regression (SSR) + Total squared error (SSE)

$$\sum_i (y_i - \bar{y}) = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

R^2 is generally between 0 and 1

Classification Error Metrics: Confusion Matrix

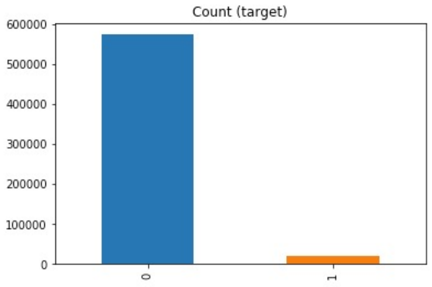
outrageously AMBITIOUS

Duke

PRATT SCHOOL *of* ENGINEERING

Accuracy

- Accuracy is the most popular and easiest to understand error metric
- However, accuracy can be deceiving when there is a class imbalance
- Consider this situation:
  - I am building a model to predict whether patients will have heart disease
  - I use data from medical study with thousands of patients and several features, along with a label of whether they were diagnosed with heart disease (“1”) or not (“0”)
  - Using this dataset, I create a classifier with 99.4% accuracy!
- What’s the problem?
  - My dataset had very high class imbalance
  - My model just predicted “0” for every patient
  - And it was right 99.4% of the time!



A Better Method – The Confusion Matrix

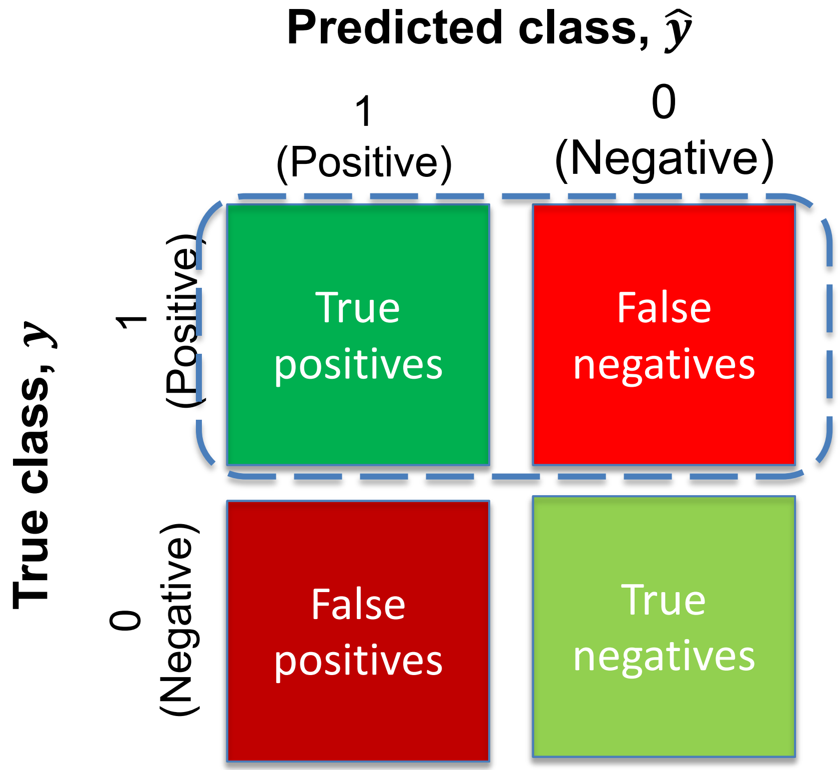
Predicted class, $\hat{y}$

True class, y	1 (Positive)	0 (Negative)
1 (Positive)		
0 (Negative)		

Confusion Matrix – Binary Classification

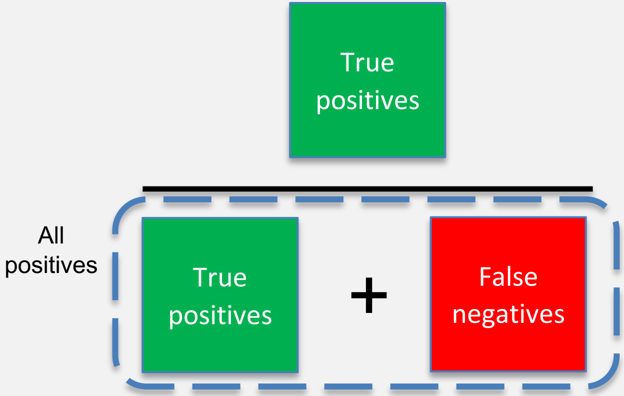
True class, y \ Predicted class, $\hat{y}$	1 (Positive)	0 (Negative)
1 (Positive)	True positives	False negatives
0 (Negative)	False positives	True negatives

Confusion Matrix – Binary Classification

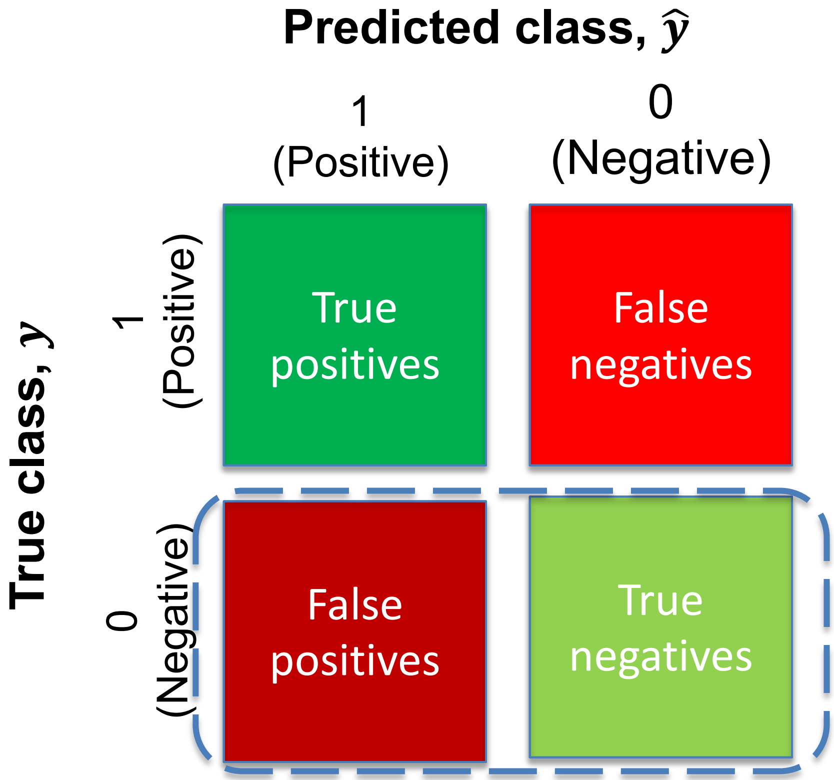


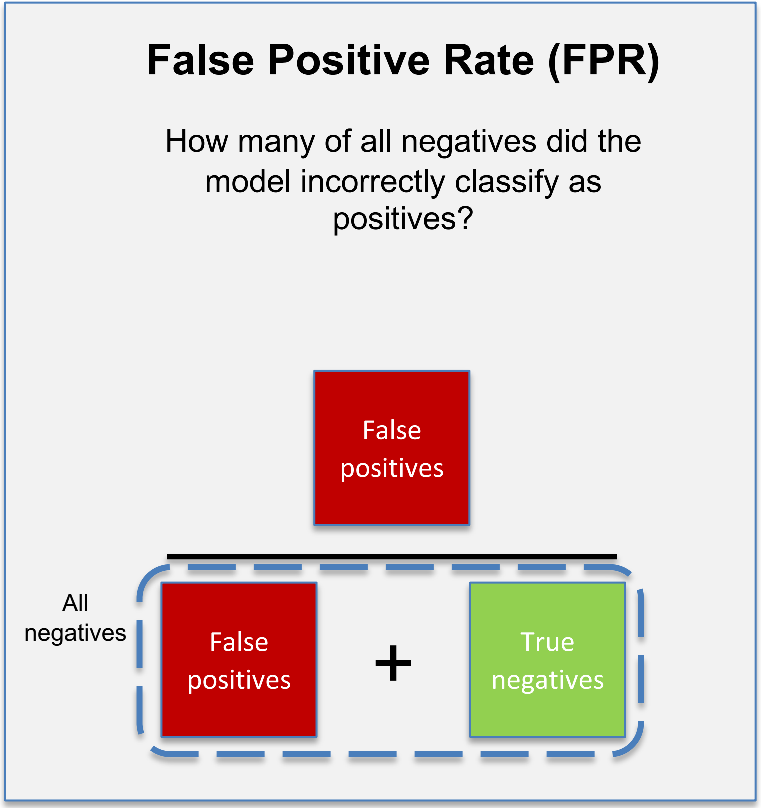
True Positive Rate (TPR)
Recall
Sensitivity

How many of all positives did the model correctly classify as positives?

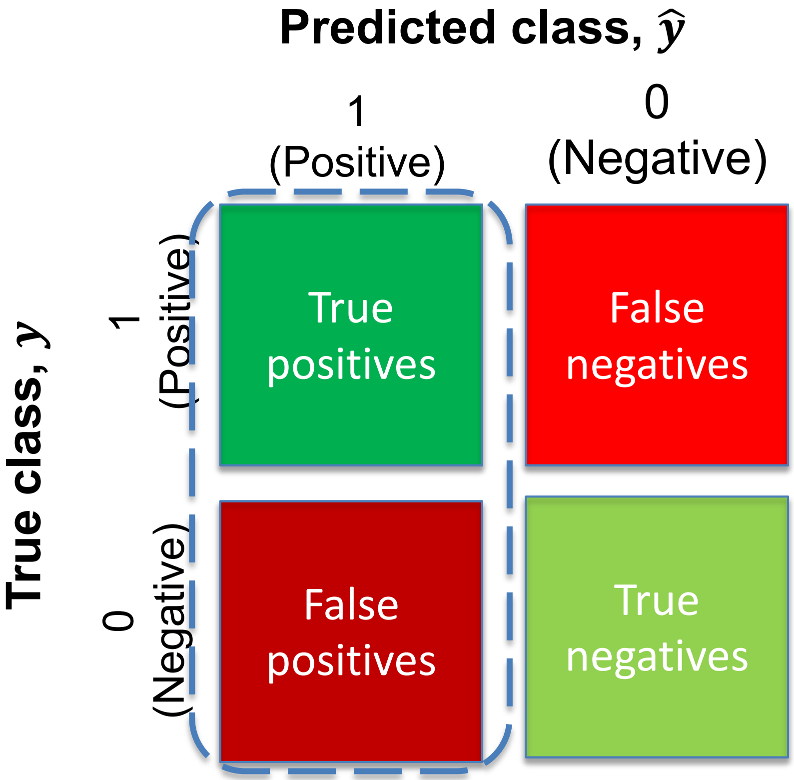


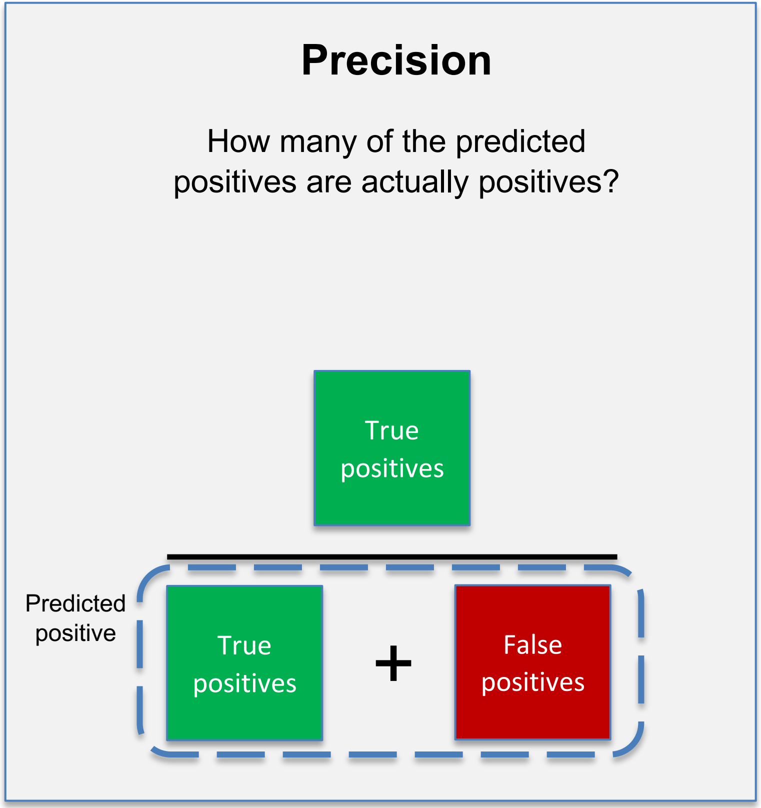
Confusion Matrix – Binary Classification





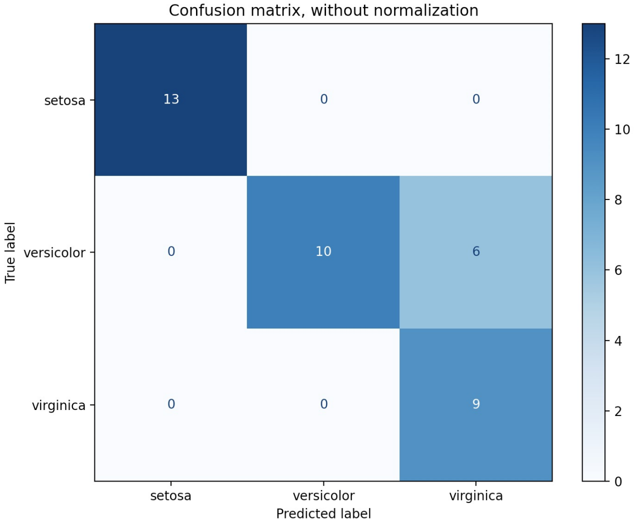
Confusion Matrix – Binary Classification





Multiclass Confusion Matrix

The multiclass confusion matrix shows us where the model struggles to differentiate between classes, and we calculate metrics per class



outrageously AMBITIOUS

Classification Error Metrics: ROC and PR Curves

Duke
PRATT SCHOOL of ENGINEERING

ROC Curves

- A **Receiver Operating Characteristic** (ROC) curve plots the *True Positive Rate (TPR)* and *False Positive Rate (FPR)* for different <u>threshold</u> values

- What is a **threshold**?
  - Most classification models return the probability of the positive class
  - We set a threshold for the positive class:

Classifier decision rule:

$$\hat{y} = \begin{cases} 1, & x > thresh \\ 0, & x \leq thresh \end{cases}$$

ROC Curves

To build a ROC curve:

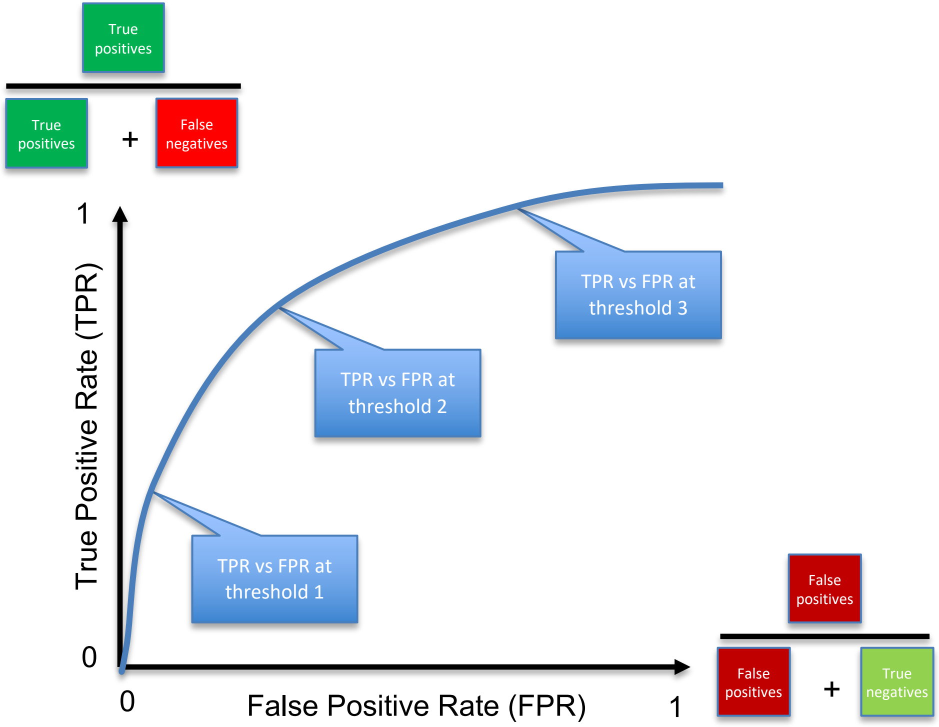
- Run the model and get the output probabilities
- For each value in range(0,1):
  - Set value as threshold value
  - Get predictions by comparing model output probabilities to threshold
  - Calculate the TPR and FPR values
- Plot the values for all thresholds on a graph of TPR vs FPR

	Target	Model Output	Thresh = 0.3	Thresh = 0.5	Thresh = 0.7
1	1	0.85	1	1	1
2	0	0.04	0	0	0
3	1	0.62	1	1	0
4	0	0.37	1	0	0
5	0	0.55	1	1	0
True Positive Rate (TPR)			2/2	2/2	1/2
False Positive Rate (FPR)			2/3	1/3	0/3

ROC Curves

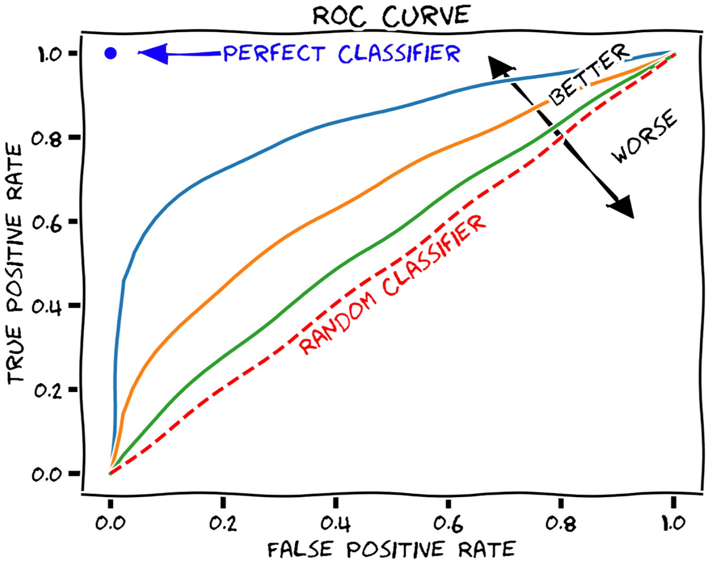
To build a ROC curve:

- Run the model and get the output probabilities
- For each value in range(0,1):
  - Set value as threshold value
  - Get predictions by comparing model output probabilities to threshold
  - Calculate the TPR and FPR values
- Plot the values for all thresholds on a graph of TPR vs FPR



Area Under ROC (AUROC)

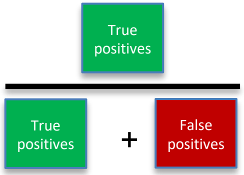
- A common error metric for classification models is the Area Under the ROC (AUROC)

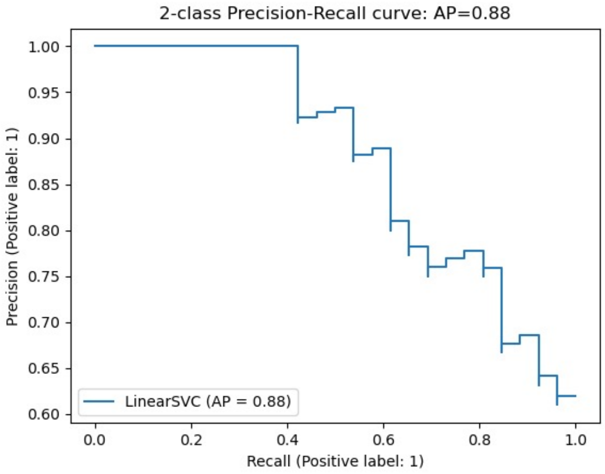


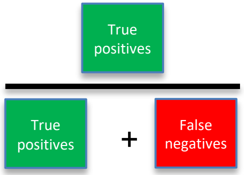
https://en.wikipedia.org/wiki/Receiver_operating_characteristic#/media/File:Roc-draft-xkcd-style.svg, CC0

Precision-Recall Curve

- Another evaluation technique is the precision-recall (PR) curve
- This measures the tradeoff between recall and precision as the model threshold is varied
- PR curves are especially useful if we have high class imbalance (e.g. a lot of 0’s and only a few 1’s)
  - Unlike ROC curves, they do not factor in True Negatives



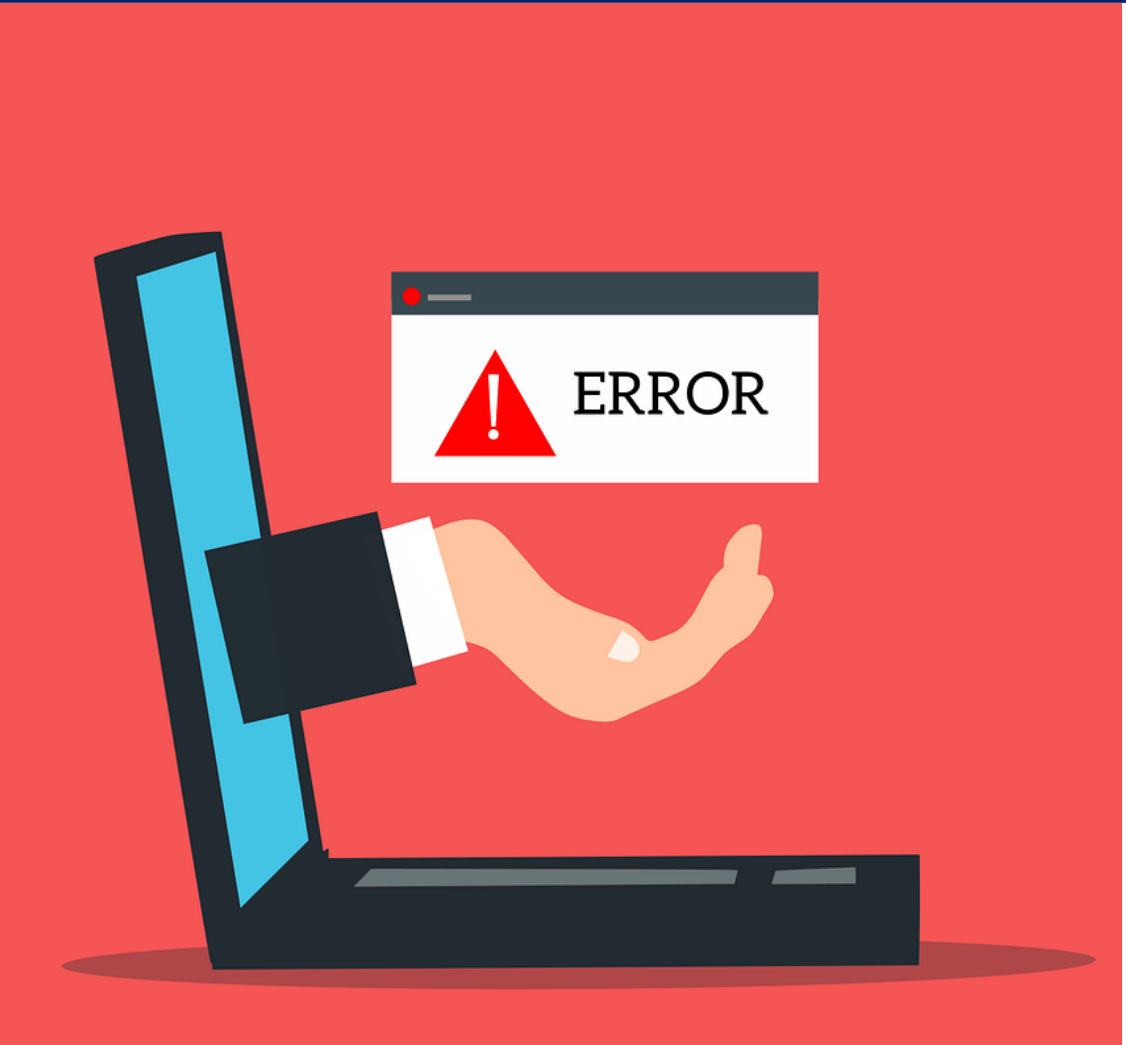




https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html#sphx-glr-auto-examples-model-selection-plot-precision-recall-py outrageously AMBITIOUS

Troubleshooting Model Performance

Duke PRATT SCHOOL of ENGINEERING


ERROR

Sources of error

1. Problem framing & metric selection
2. Data quantity & quality
3. Feature selection
4. Model fit
5. Inherent error

Sources of error

1. **Problem framing & metric selection**
2. Data quantity & quality
3. Feature selection
4. Model fit
5. Inherent error

Sources of error

1. Problem framing & metric selection
2. **Data quantity & quality**
3. Feature selection
4. Model fit
5. Inherent error

Sources of error

1. Problem framing & metric selection
2. Data quantity & quality
3. **Feature selection**
4. Model fit
5. Inherent error

Sources of error

1. Problem framing & metric selection
2. Data quantity & quality
3. Feature selection
4. **Model fit**
5. Inherent error

Sources of error

1. Problem framing & metric selection
2. Data quantity & quality
3. Feature selection
4. Model fit
5. **Inherent error**

outrageously
AMBITIOUS

Wrap-Up

Duke
PRATT SCHOOL *of* ENGINEERING

Wrap-Up: Metric Selection

- Selecting proper <u>outcome</u> and <u>output</u> metrics is key to a successful machine learning project
- Your choice of metric should reflect the nature of your problem and the consequences of being wrong
  - For a regression problem, is it worse to be very wrong a few times, or a little wrong a lot of times?
  - For a classification problem, are false positives or false negatives worse?