Predictive Modeling

Danny Lumian, Ph.D.
Data Science Training Specialist
Office of Data Science Strategy
National Institutes of Health

Types of Targets

- Classification
  - Target to predict is one of a set of classes
  - Examples:
    - Spam or not spam
    - Type of a flower
    - Malignant or benign tumor
- Regression
  - Target to predict is numeric/continuous quantity
  - Examples:
    - Life expectancy
    - Cost of housing
    - Miles per gallon of a car

Three types of variables

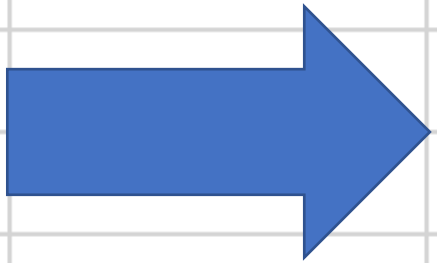
- Independent Variable
  - Predictive variable(s)
  - May be manipulated by researcher, such as conditions in an experiment
- Dependent Variable
  - The outcome or target (y) variable of interest
- Controlled Variable
  - Variable that may impact the outcome or dependent variable
  - Measured to help ensure rigorous standards for an experiment

Train-Test Split

- It is common in machine learning to create a train-test split
- The train data is used to train the model
  - This data is “seen” by the model
- The test data is used to evaluate model performance
  - This data is “unseen” by the model
- Test metrics are usually more informative than train metrics as it shows how well the model generalizes its predictions
- Common splits are 80/20 or 70/30
- Can be useful to stratify data based on target

Dummy Coding Variables

Group Assignment		Group 2	Group 3
Group 1		0	0
Group 2		1	0
Group 1		0	0
Group 3		0	1
Group 2		1	0



Convert categorical variables to a set of binary variables
Drop 1 group as the baseline
All other groups get a new numeric column

Bias and Variance

- The *bias* error is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
- The *variance* is an error from sensitivity to small fluctuations in the training set. High variance may result from an algorithm modeling the random noise in the training data (overfitting).

From Wikipedia: https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff

Underfitting vs Overfitting

Underfitting

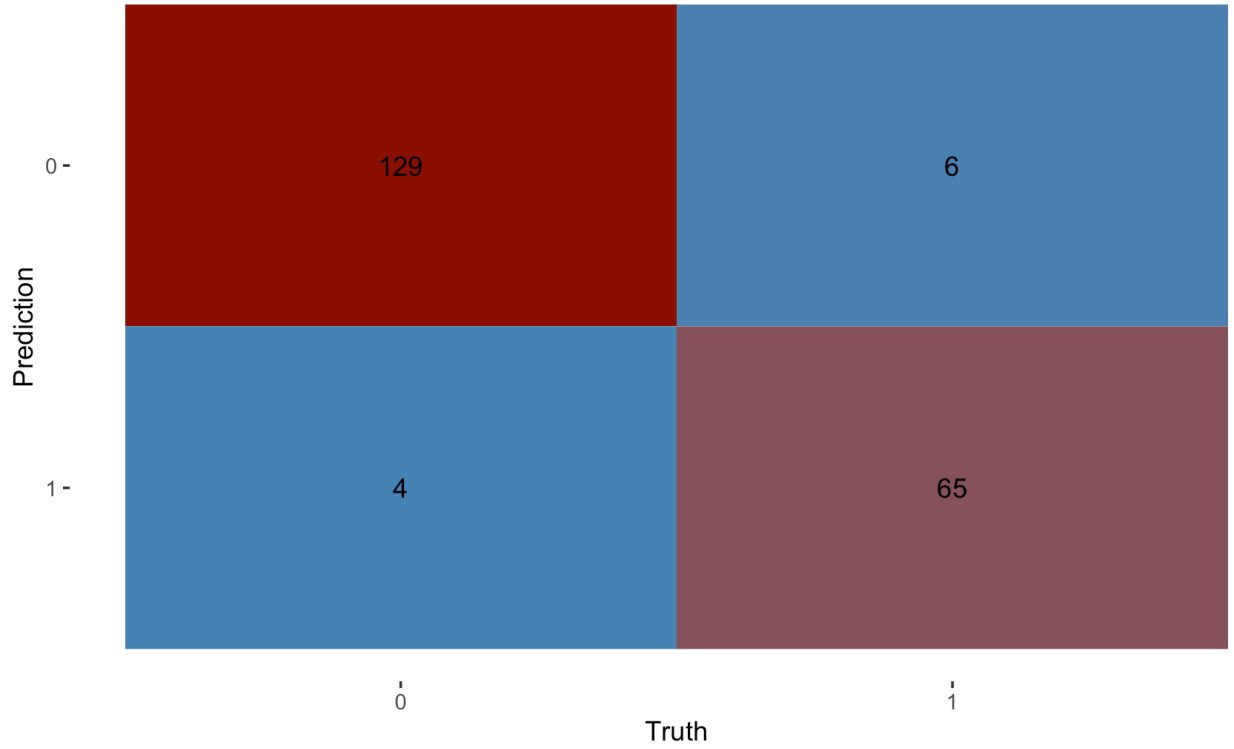
- A model that does not capture the nuance of the data
- Example might be a model that just predicts the mean or dominant class
- Can add features or try a different model

Overfitting

- A model that does not generalize well
- Attends to meaningless patterns in the data
- Detected by good training metrics but poor test metrics
- Can remove features or tune the model

Classification Outcomes

- True Negatives (TN)
  - Top left quadrant
  - Was not a positive case and was not predicted as a positive case
- False Negatives (FN)
  - Top right quadrant
  - Was a positive case but was predicted as a negative case
- True Positives (TP)
  - Bottom left quadrant
  - Was a positive case and was predicted as a positive case
- False Positives (FP)
  - Bottom left quadrant
  - Was a negative case but was predicted as a positive case

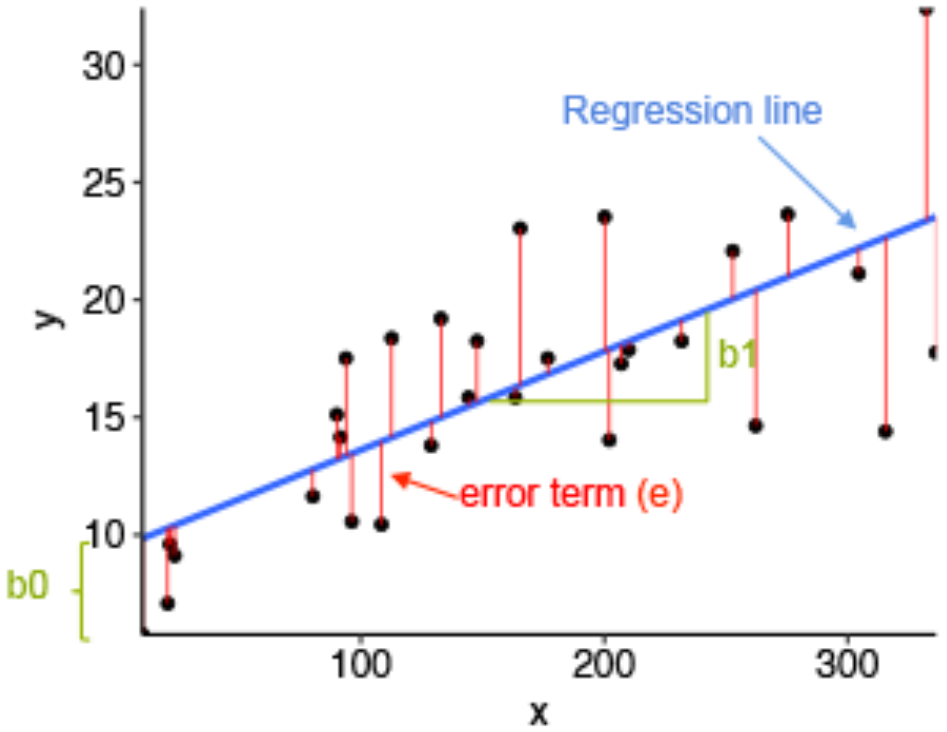


Classification Metrics

- Accuracy (TP+TN)/Total
  - Percentage of total predictions that were correct
- Sensitivity (TP/(TP+FN))
  - True positive rate
- Specificity (TN/(TN+FP))
  - True negative rate
- Precision (TP/(TP+FP))
  - Positive prediction value
- F1-Score (2TP/(2TP+FP+FN))
  - Harmonic mean of precisions and sensitivity

Regression Metrics

Mean squared error	$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$
Mean absolute percentage error	$\mathrm{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n} \left\lvert \frac{e_t}{y_t} \right\rvert$



Missing Data

- Many models can not handle missing data
- There are several solutions to address this:
  - Remove missing data rows
  - Impute missing data
    - Example: fill with the mean or most common category
  - More sophisticated imputation methods exist

Imbalanced Data Sets

- Undersample
  - Remove rows of the more common class
- Oversample
  - Replicate the minority class
- Synthetic Data
  - Create fake data based on observations and use those data in the model

Questions?