
CLASSIFICATION
MODELS
Paul Speaker

# Classification Models

- Classification models are models which seek to use various input variables to predict a categorical target
- Like with continuous targets, there are two main viewpoints to view classification models
  - Probabilistic: the model gives a probability of being in each of the different possible categories, based on maximum likelihood
  - Least Loss: the model is such that it minimizes the error (measured in different ways) for the predictions
- Naïve classification model: no inputs, prediction is always most common target value
  - Error rate is less than or equal to 50%
  - Comparisons of model should always be vs. baseline

# Types of Classification Models

- Classical Methods
  - Classical methods use data transformations to turn a classification model into a continuous model
    - Logistic Regression
    - Discriminant Models
- Tree-Based methods
  - Tree-based methods use discrete rules to find the best way to classify targets
    - Decision trees
    - Random forests
    - xgboost
- Neural Networks
  - Neural networks are similar to either, but use “hidden layers” to create nonlinear composition of activation functions
- Non-parametric methods
  - Non-parametric methods use local properties to make classifications
  - K-nearest neighbors (KNN)

# Probabilistic Interpretation of Classification Models

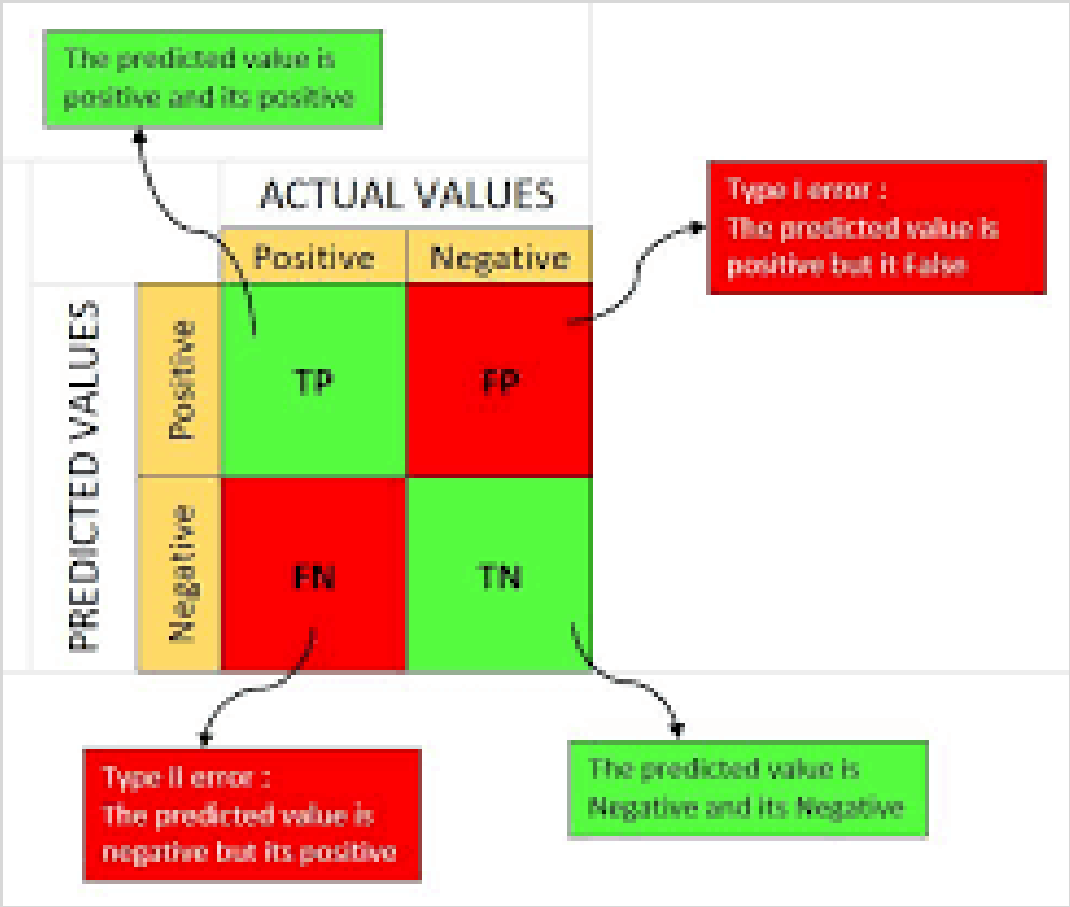
- Any classification model can be interpreted as a probability of being each categorical value
- It is common to reduce the probability to classification assignment, but this can be a mistake
  - If you care about aggregates, it can cause problems. For example, if 5 instances each have a 40% chance of being a target value, assigning a classification target will give 0 instances in the aggregate, even through the expected number will be 2 (0.4 x 5)
- Predicted Probabilities: The model will output a probability for each class, representing the likelihood that a given instance belongs to that class. These probabilities can be used to make decisions about which class to predict, depending on the desired level of confidence.
- Calibration: It measures the consistency between predicted probabilities and the true outcomes. A well-calibrated model will have predicted probabilities that align well with the actual outcomes.

# EDA for Classification Models

- EDA for classification models shares several characteristics with EDA for continuous target models
- First target for EDA is the target
  - Plot a class histogram for the target
  - Will want to know whether target is "balanced"
    - Balanced = all classes roughly the same counts
      - 60/40 is still balanced, 90/10 is not
- Next, it is helpful to look at the relationship between the target and possible inputs
- Will be different for numerical vs. categorical X's
- Faceting by the classification result is a great way to handle this

# Tools to Assess Classification Models

- Confusion Matrix



# Tools to Assess Classification Models

- ROC Curves: assess true/false positives at different positivity thresholds