Mar 13, 2023

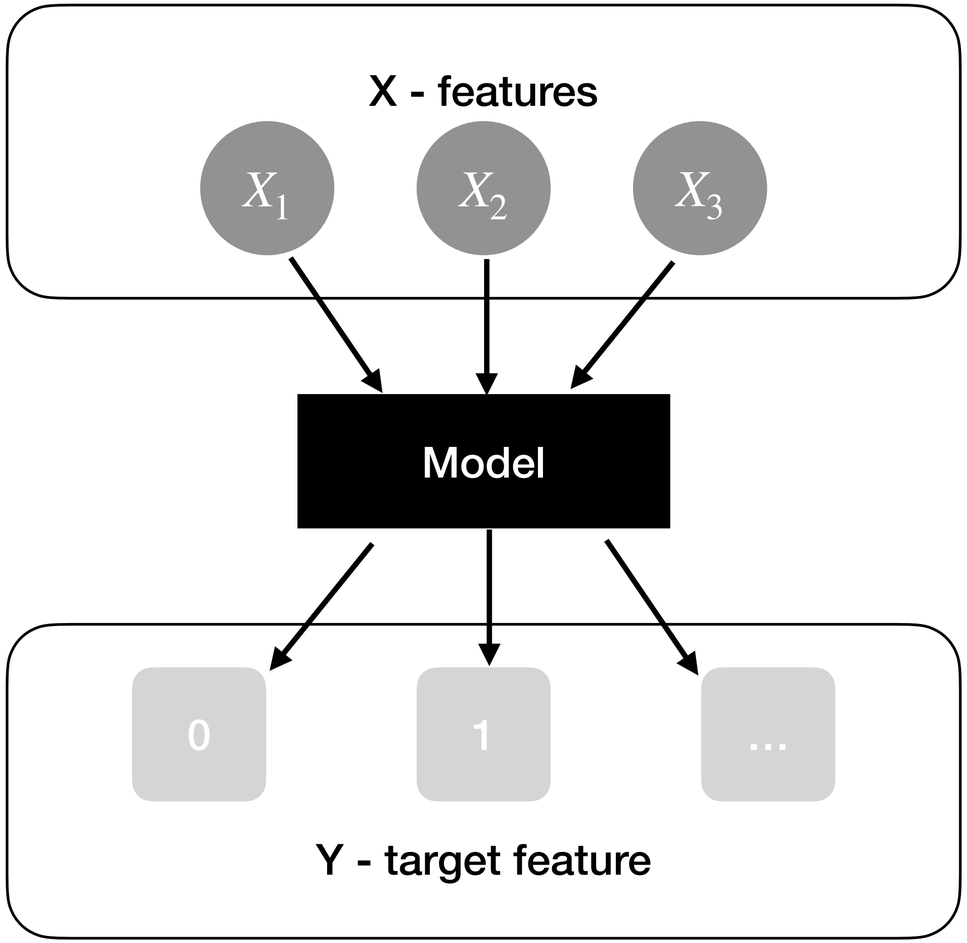
Machine Learning Methods and Applications

Week 3. Supervised Learning: Logistic Regression Models

Remember

- Data splitting is an important step to see of how well the model performs on unseen (test) data.
- Use strategies to avoid that a model should not be suffer from overfitting and underfitting problem.
- Bias-variance tradeoff

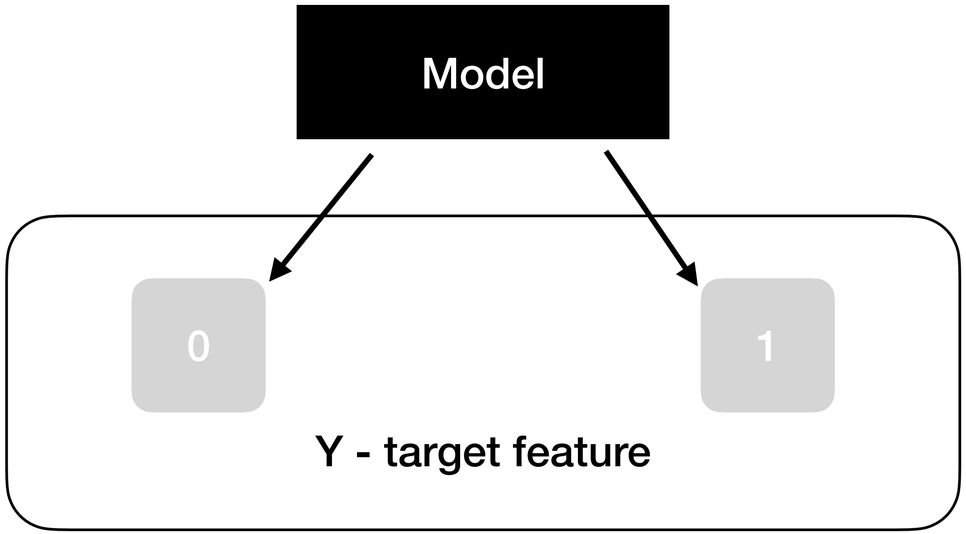
Classification task



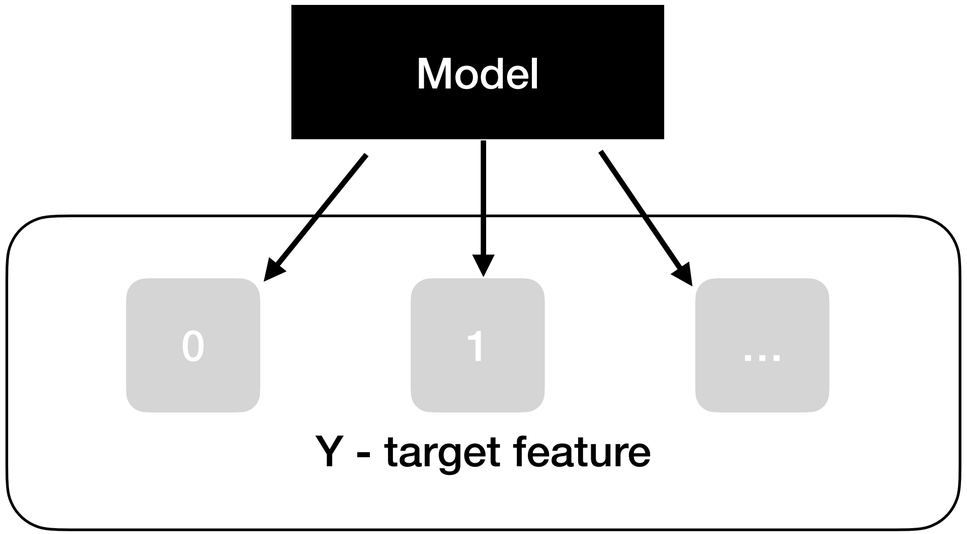
- E-mail spam detection (spam or not)
- Churn prediction (churn or nut)
- Conversion prediction (buy or not)

Type of classification tasks

Binary classification
(two classes target feature)



Multi-class classification
(more than two classes target feature)



Why not linear regression models?

Linear regression is used to approximate the linear relationship between a continuous target feature and set of features.

Logistic regression model

Simple LRM: one feature

$$Y = \beta_0 + \beta_1 X$$

Multiple LRM: multiple features

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Xs: features, βs: parameters, Y: categorical target

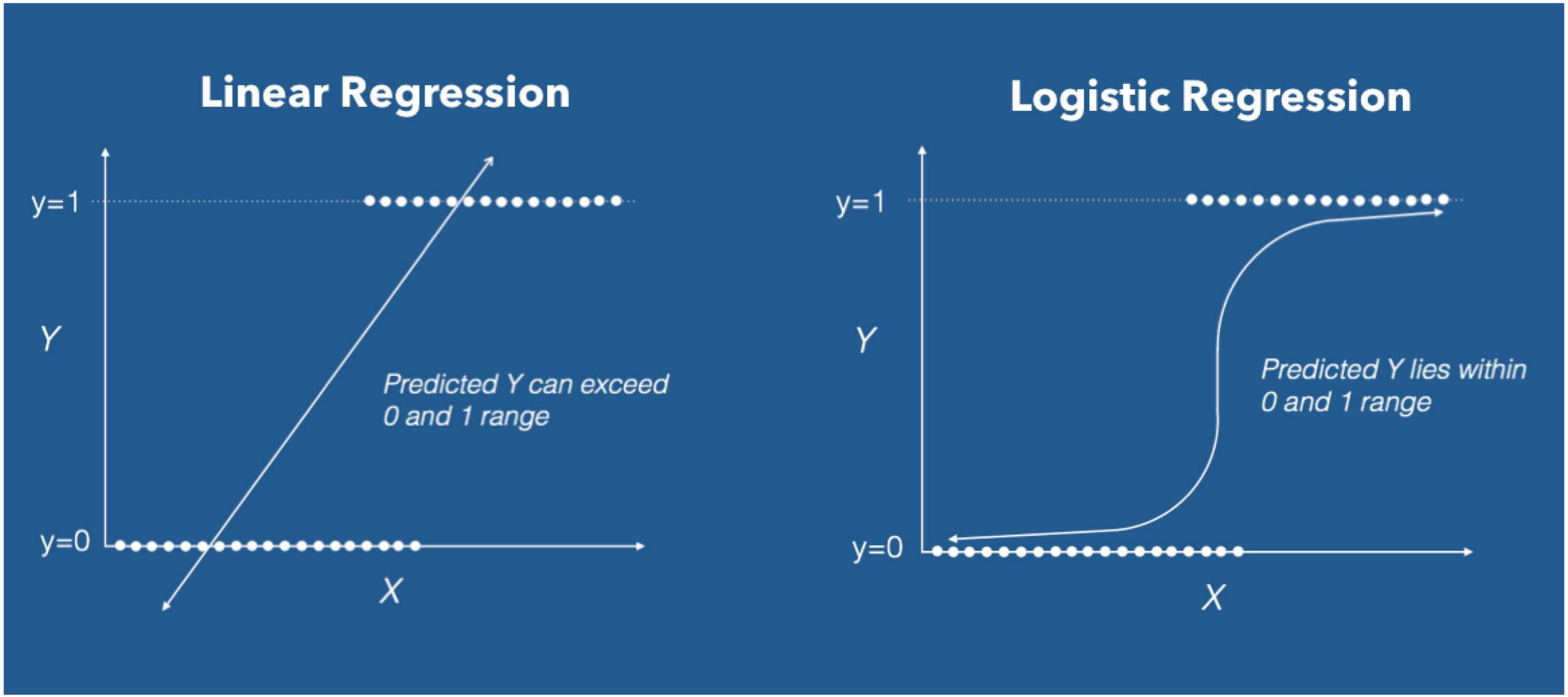
Logistic regression model

Let Y be a binary/dichotomous target variable takes the values $\{0,1\}$. p is the probability of $y = 1$, $p = P(Y = 1)$. The binary logistic regression models is

$$logit(p) = log(p/1 - p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

- The transformation from probability to odds is a monotonic transformation, meaning the odds increase as the probability increases or vice-versa.
- Probability (p) ranges from 0 and 1.
- Odds ratio $(p/1 - p)$ ranges from 0 and ∞.

Linear regression vs. Logistic regression



Steps of classification models

Steps

1. Data splitting
2. Model training
3. **Measuring model performance**

Step 1. Data splitting

Missing (NA) values in observations

In the step of splitting data set, some of the classes of a categorical feature may not be assign to the train set, then the model can not learn anything about these classes. Thus, it is not possible to predict the value of target feature with the predictors (features) that have these unseen classes.

Thus, it is necessary to handle the observations that have missing values.

Step 1. Data splitting

Missing (NA) values in observations

It is necessary to handle the observations that have missing values. The solution ways:

- Removing (may lead loss of information,
  easier way if the data set is enough large*)
- Imputation (may lead overfitting or bias)

*no consensus about the size of enough large dataset, it depends on the domain.

Step 3. Measuring the model performance

Confusion matrix		**Observed class**	
		Positive (1)	Negative (0)
Predicted class	Positive (1)	True Positive (TP)	False Positive (FP)
	Negative (0)	False Negative (FN)	True Negative (TN)

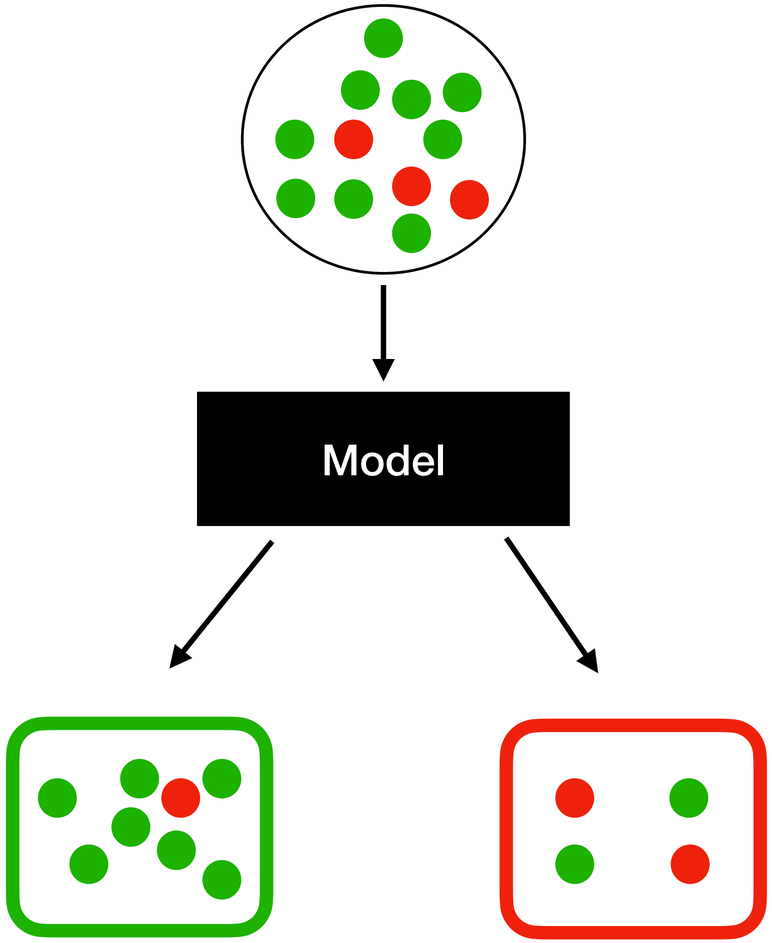
$$\text{Sensitivity (Recall)} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Accuracy} = \frac{TN + TP}{TP + FP + TN + FN}$$

Step 3. Measuring the model performance



Confusion matrix		Observed class	
		🟢	🔴
Predicted class	🟢	TP = 7	FP = 1
	🔴	FN = 2	TN = 2

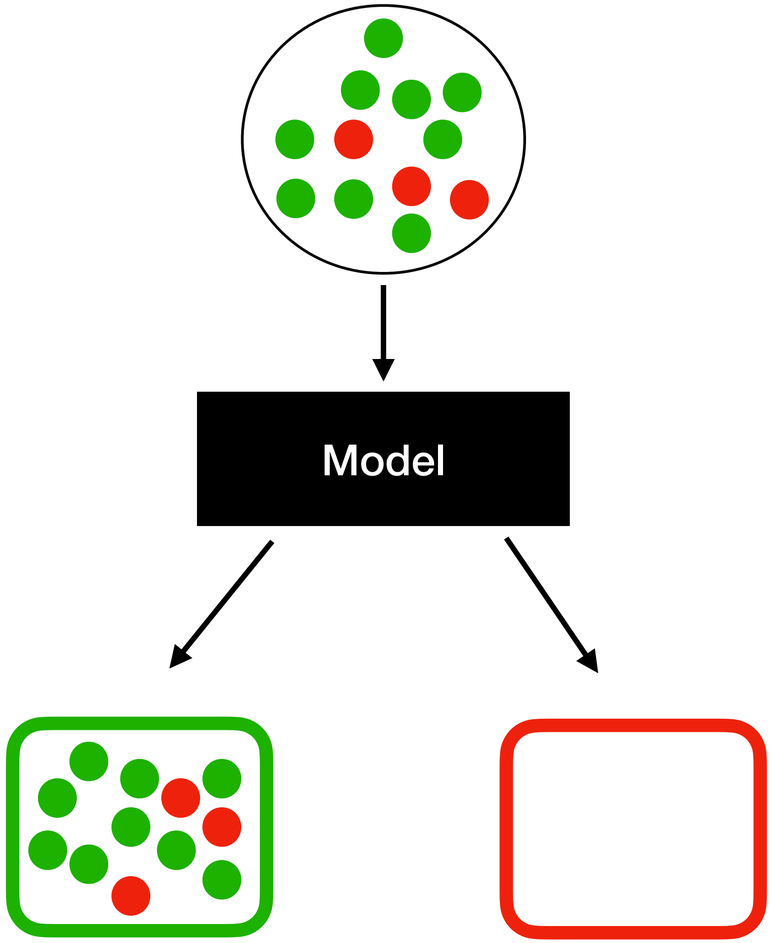
Sensitivity (Recall) $= \frac{TP}{TP + FN} = \frac{7}{7 + 2} = 0.78$

Specificity $= \frac{TN}{TN + FP} = \frac{2}{2 + 1} = 0.67$

Precision $= \frac{TP}{TP + FP} = \frac{7}{7 + 1} = 0.88$

Accuracy $= \frac{TN + TP}{TP + FP + TN + FN} = \frac{2 + 7}{7 + 1 + 2 + 2} = 0.75$

Step 3. Measuring the model performance



Confusion matrix		Observed class	
		●	●
Predicted class	●	TP = 9	FP = 3
	●	FN = 0	TN = 0

$$\text{Sensitivity (Recall)} = \frac{TP}{TP + FN} = \frac{9}{9 + 0} = 1$$

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{0}{0 + 3} = 0$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{9}{9 + 0} = 1$$

$$\text{Accuracy} = \frac{TN + TP}{TP + FP + TN + FN} = \frac{0 + 9}{9 + 3 + 0 + 0} = 0.75$$

Accuracy may not show the overall performance of the model

Step 3. Measuring the model performance

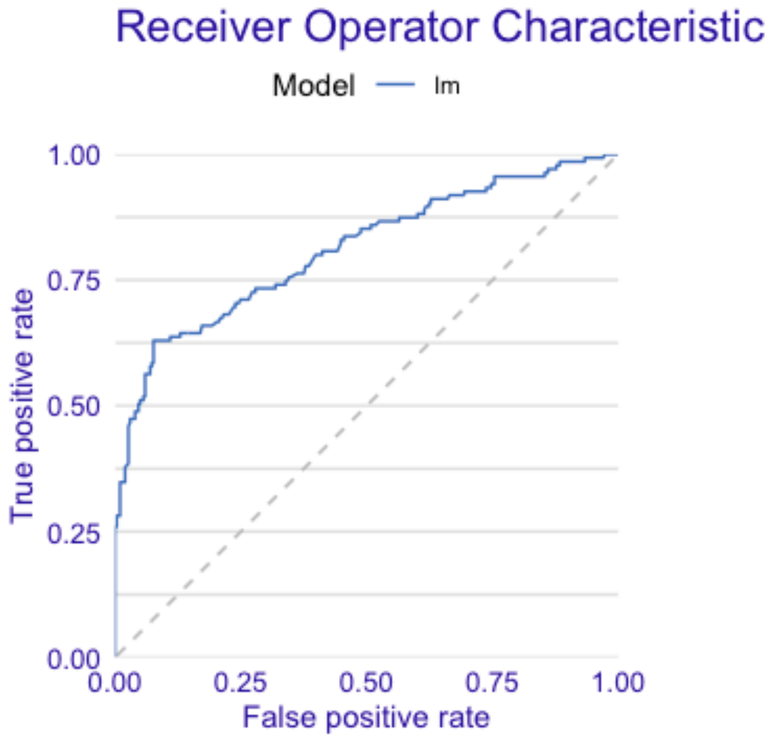
The limitation of the metrics:

- Any model performance metrics in classification is depend on the cut-off value (c) to assign a class to an observation.
- In general, the model prediction $p > c$ then the class is assigned as 1, or 0 otherwise.
- Thus, the metrics are depend on the value of c. Other metrics may be used in that case.

Step 3. Measuring the model performance

ROC Curve

All the metrics depend on the choice of cut-off value. To assess the form and the strength of dependence, a common way is to construct the **Receiver Operating Characteristics (ROC) curve**. The curve plots the Recall in the function of 1-Specificity for all possible values of cut-off.

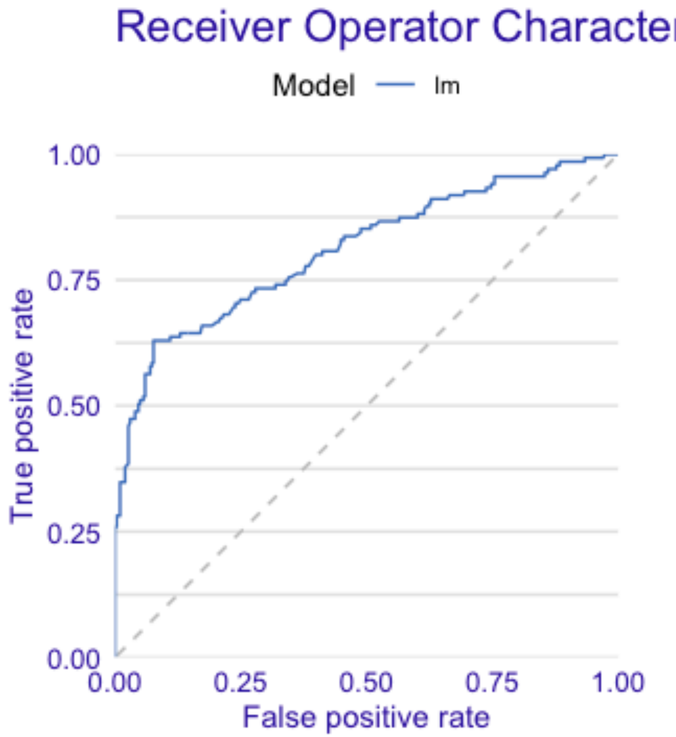


Step 3. Measuring the model performance

Area Under the Curve

To use the ROC curve as a metric, it is need to summarize the curve in a way. **Area Under the Curve (AUC)** is used as a statistics to show the summarization of ROC curve.

The higher AUC indicates higher performance!



AUC = 0.81

Application

See the R codes on the course GitHub repository!

The video recording of today's lecture will be available on YouTube, and slides on **GitHub**.
Feel free to contact me via e-mail: mustafacavus@eskisehir.edu.tr