MGS 3701 / MKT 3950: Data Mining

Chapter 5: Evaluating Predictive Performance

Chungil Chae

温州肯恩大学
WENZHOU-KEAN UNIVERSITY

2022/01/01 (updated: 2022-05-05)

Chapter Objectives

How the predictive performance of data mining methods can be assessed?

- the danger of overfitting to the training data
- the need to test model performance on data that were not used in the training step
- popular performance metrics.
  - Average Error, MAPE, RMSE (based on the validation data)
- For classification tasks (metrics based on the confusion matrix)
  - overall accuracy, specificity, sensitivity, misclassification costs.
- the relation between the choice of cutoff value and classification performance
- the ROC curve, which is a popular chart for assessing method performance at different cutoff values
- lift charts
- the need for oversampling rare classes and how to adjust performance metrics for the oversampling.
- the usefulness of comparing metrics

5.1 Introduction

Why Evaluate?

- Multiple methods are available to classify or predict
- For each method, multiple choices are avaiable for settings
- To choose best model, need to assess each model's performance

Three main types of outcomes of interest are:

- Predicted numerical value: when the outcome variable is numerical (e.g., house price)
- Predicted class membership: when the outcome variable is categorical (e.g., buyer/nonbuyer)
- Propensity: the probability of class membership, when the outcome variable is categorical (e.g., the propensity to default)

Predeiction method vs. Classification methods

- Prediction methods -> generating numerical predictions
- classification methods (“classifiers”) -> generating propensities (predicted class memberships)

Two distinct predictive uses of classifiers:

- **classification**, is aimed at predicting class membership for new records.
- **ranking** is detecting among a set of new records the ones most likely to belong to a class of interest

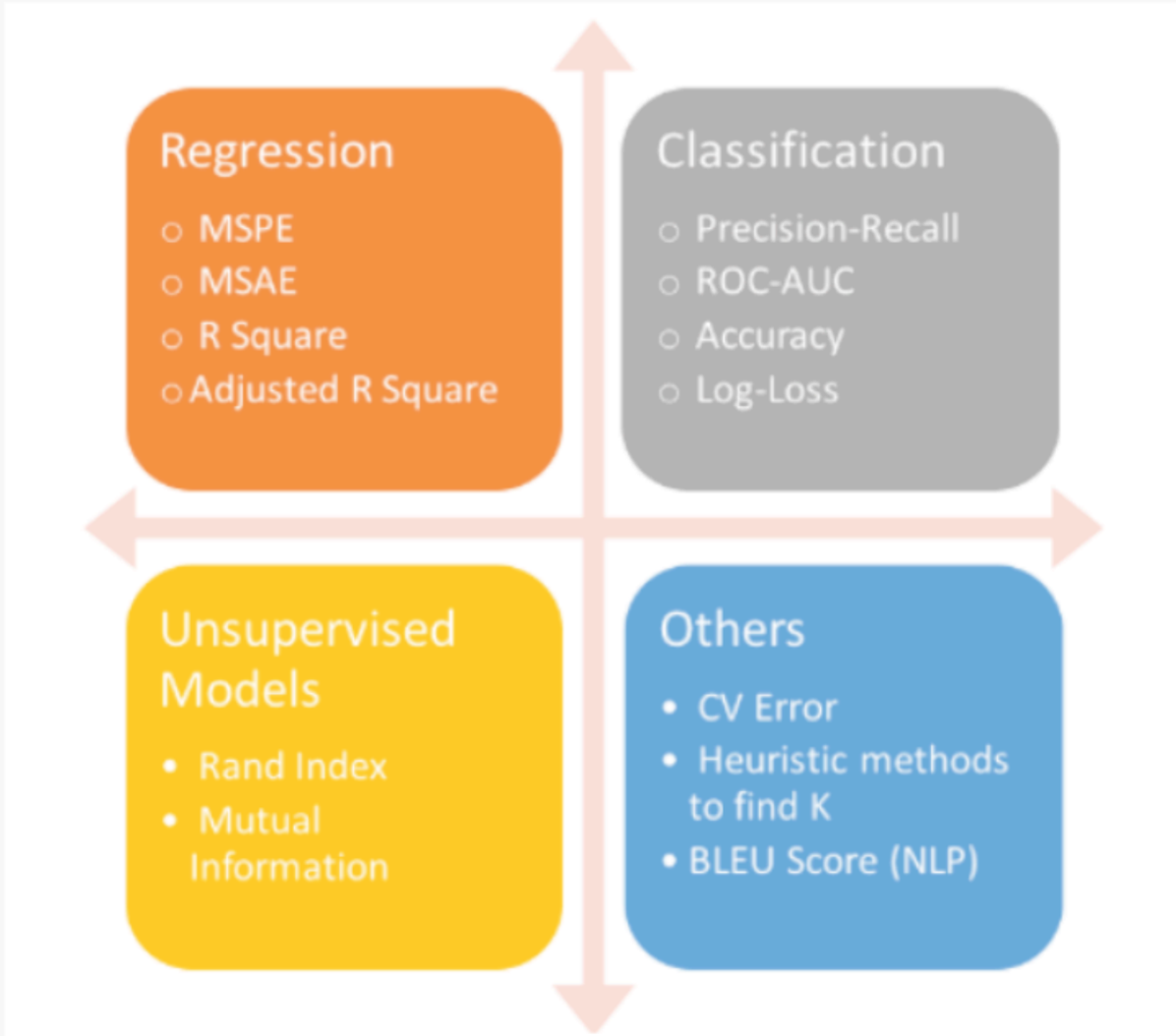
5.2 Evaluating Predictive Performance

What we want

- Not the same as "goodness-of-fit"
- We want to know how well the model predicts new data, not how well it fits the data it was trained with
- Key component of most measures is difference between actual y and predicted y (“error”)

For assessing prediction performance

- The measures are based on the validation set
- Procedure
  - Models are trained on the training data
  - Applied to the validation data
  - Measure accuracy
  - Use the prediction errors on that validation set


Regression
MSPE
MSAE
R Square
Adjusted R Square
Classification
Precision-Recall
ROC-AUC
Accuracy
Log-Loss
Unsupervised Models
Rand Index
Mutual Information
Others
CV Error
Heuristic methods to find K
BLEU Score (NLP)

Naive Benchmark: The Average

- The benchmark criterion in prediction is using the average outcome value
- the prediction for a new record is simply the average across the outcome values of the records in the training set (y)
- A good predictive model should outperform the benchmark criterion in terms of predictive accuracy

So called

- Baseline

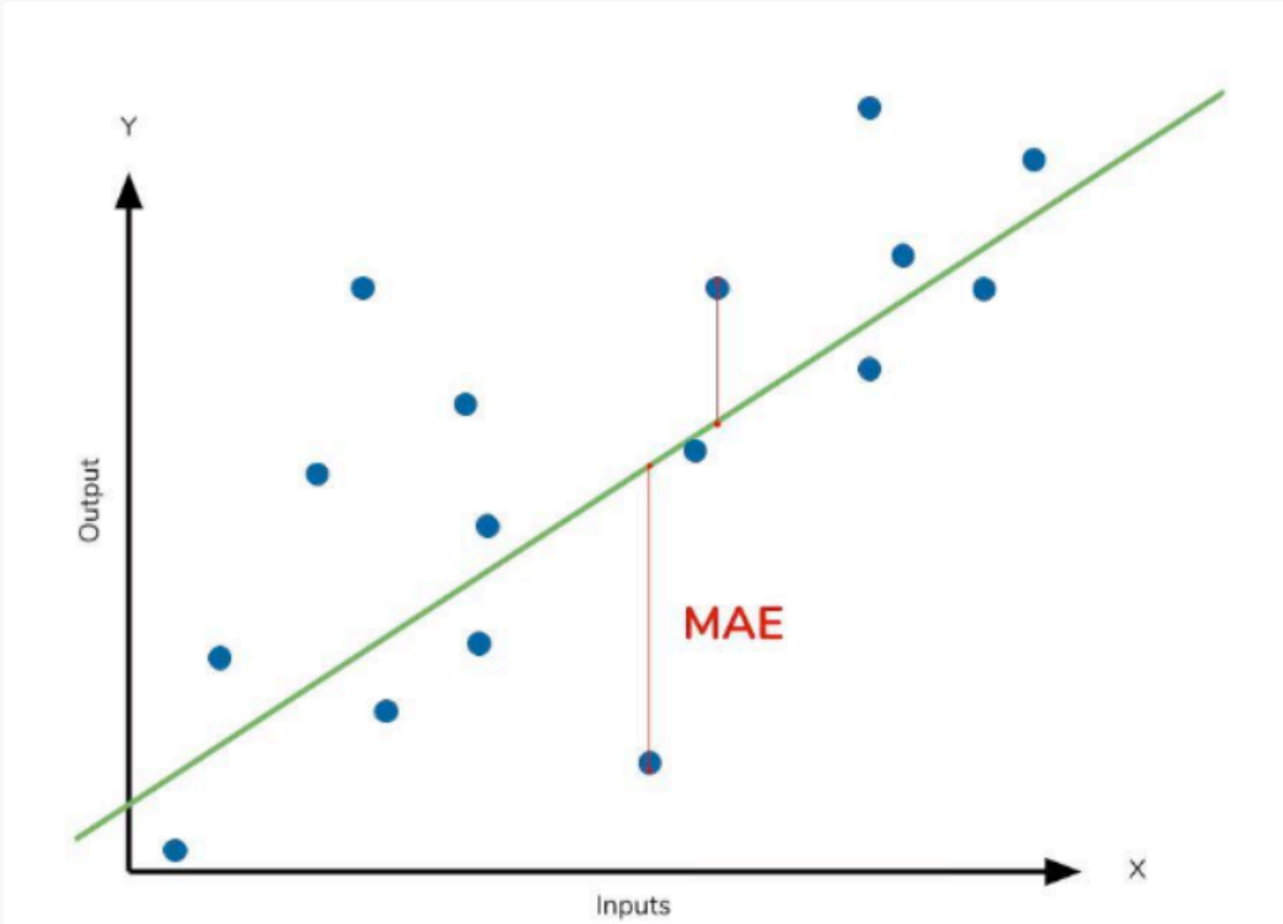
Prediction Accuracy Measures

- **Mean Error (ME)**
- **Mean Absolute Error/deviation (MAE)**
- **Mean Percentage Error (MPE)**
- **Mean Absolute Percentage Error (MAPE)**
- **Root Mean Squared Error (RMSE)**

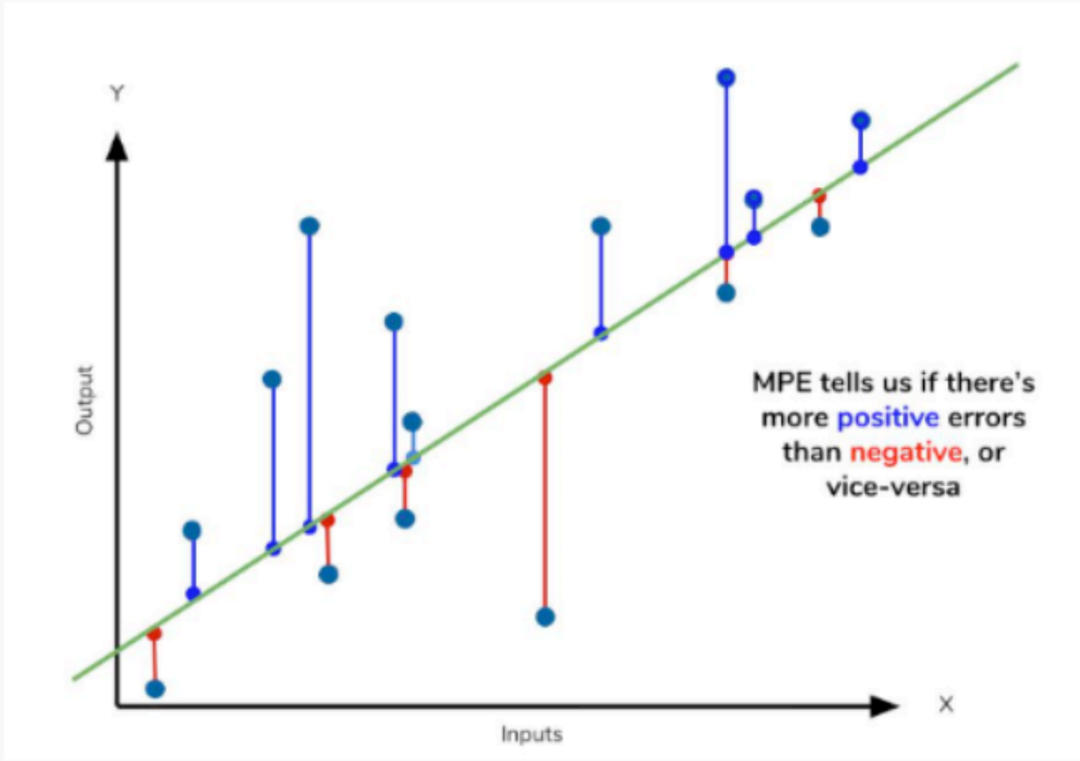
$$e = y - \hat{y}$$

- **Mean Error (ME)**
  - $\frac{1}{n}\sum_{i=1}^{n} e_i$
  - Whether the predictions are on average over- or underpredicting the outcome variable.
  - Disadvantages that show robust results against outliers and are less sensitive to errors compare to MSE,RMSE
  - The error value is relatively less affected by outliers
  - Dependent on the scale

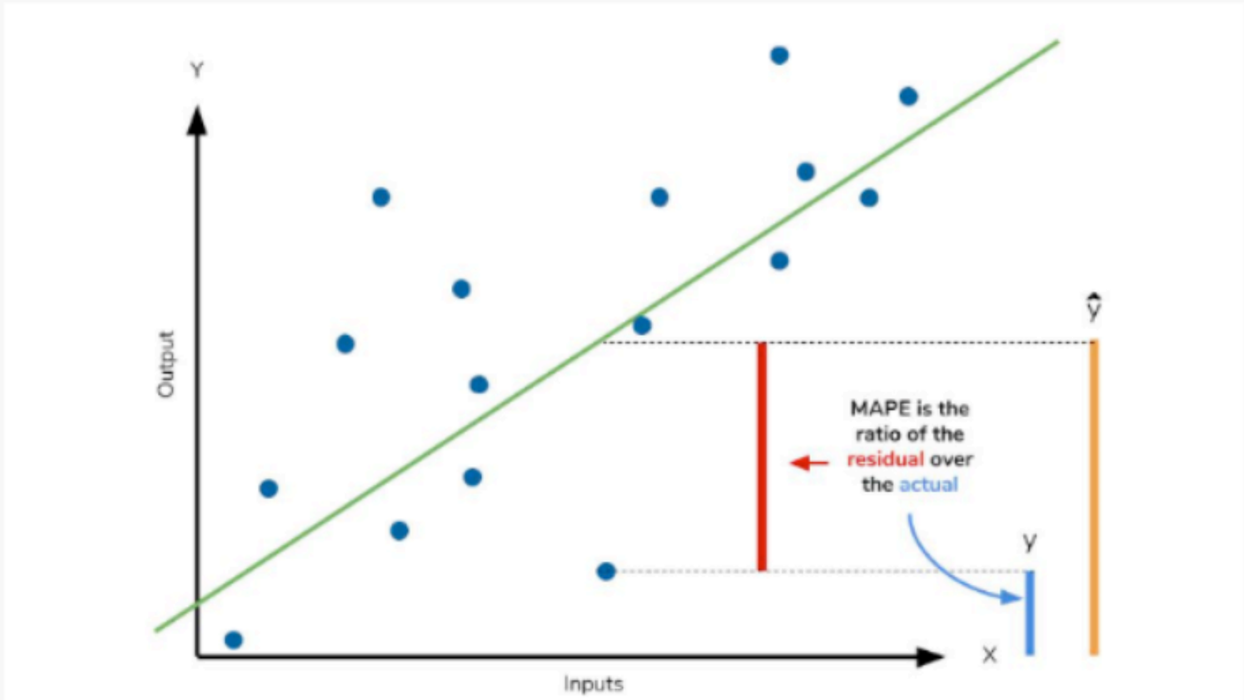
- **Mean Absolute Error/deviation (MAE)**
  - $\frac{1}{n}\sum_{i=1}^{n}|e_i|$
  - The magnitude of the average absolute error



- **Mean Percentage Error (MPE)**
  - $100 * \frac{1}{n} \sum_{i=1}^{n} e_i / y_i$
  - the percentage score of how predictions deviate from the actual values (on average), taking into account the direction of the error



- **Mean Absolute Percentage Error (MAPE)**
  - $100 * \frac{1}{n} \sum_{i=1}^{n} |e_i / y_i|$
  - a percentage score of how predictions deviate (on average) from the actual values



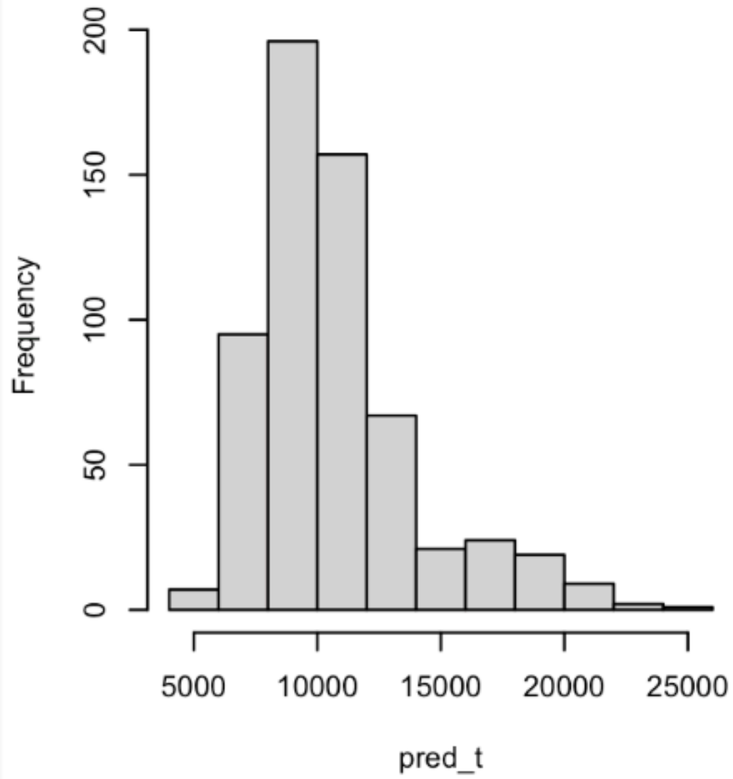
ŷ is smaller than the actual value	ŷ is greater than the actual value
n = 1 ŷ = 10 y = 20	n = 1 ŷ = 20 y = 10
MAPE = 50%	MAPE = 100%

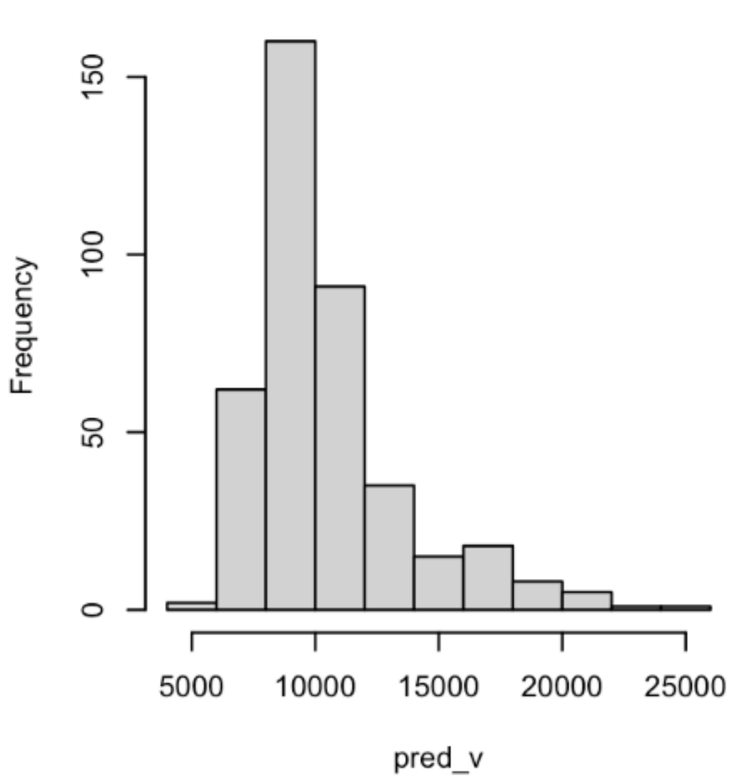
- **Root Mean Squared Error (RMSE)**
  - $\sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$
  - This is similar to the standard error of estimate in linear regression
    - the average distance that the observed values fall from the regression line
    - It tells you how wrong the regression model is on average using the units of the response variable.

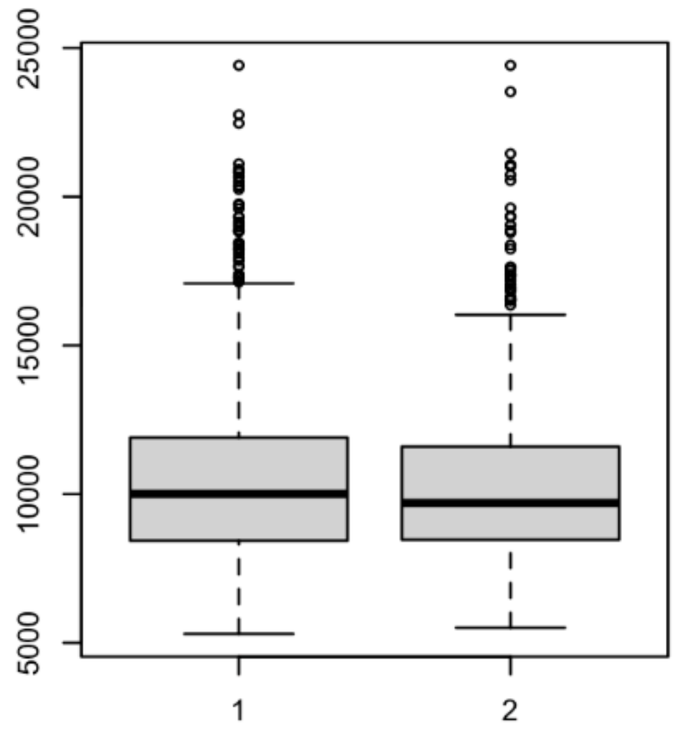
> If you want more, refer to the following link: https://arxiv.org/abs/1809.03006

- These are the result of fitting a certain predictive model to prices of used Toyota Corolla cars.
- The training set includes 600 cars and the validation set includes 400 cars.
- Results are displayed separately for the training and validation sets.

```
##                      ME    RMSE     MAE   MPE  MAPE
## training set       0.00 1036.53 787.34 -0.96  8.00
## validation set     4.92 1035.49 786.07 -1.27  8.16
## difference        -4.92    1.04   1.27  0.31 -0.16
```


Histogram of pred_t
Frequency
0
50
100
150
200
5000
10000
15000
20000
25000
pred_t


Histogram of pred_v
Frequency
0
50
100
150
5000
10000
15000
20000
25000
pred_v

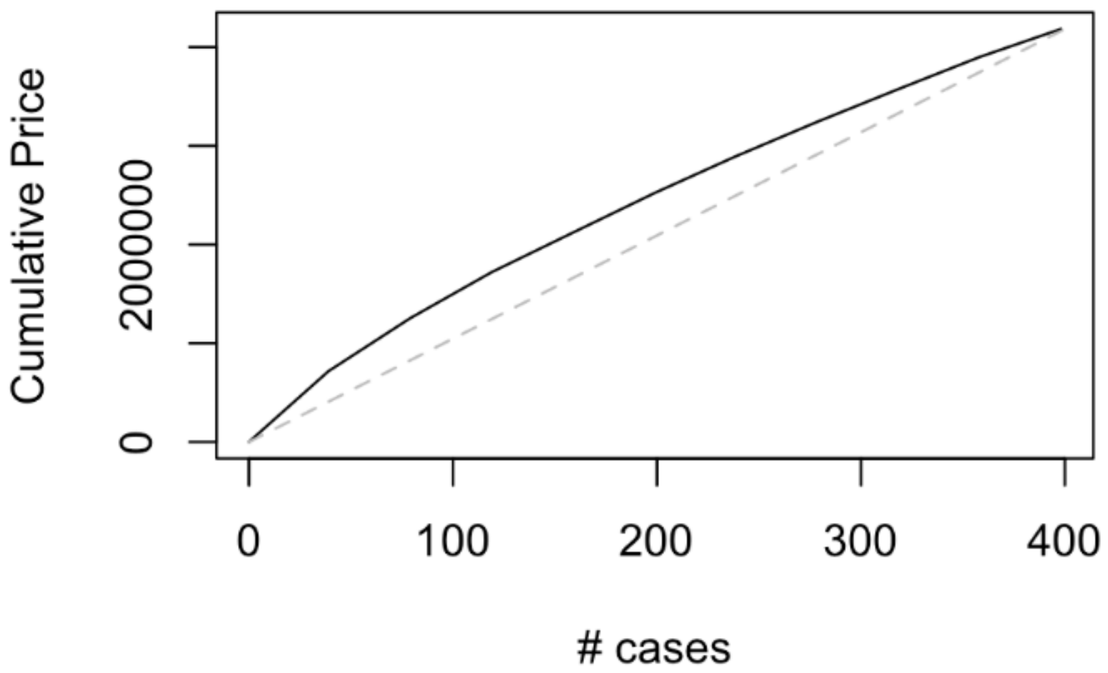

5000
10000
15000
20000
25000
1
2

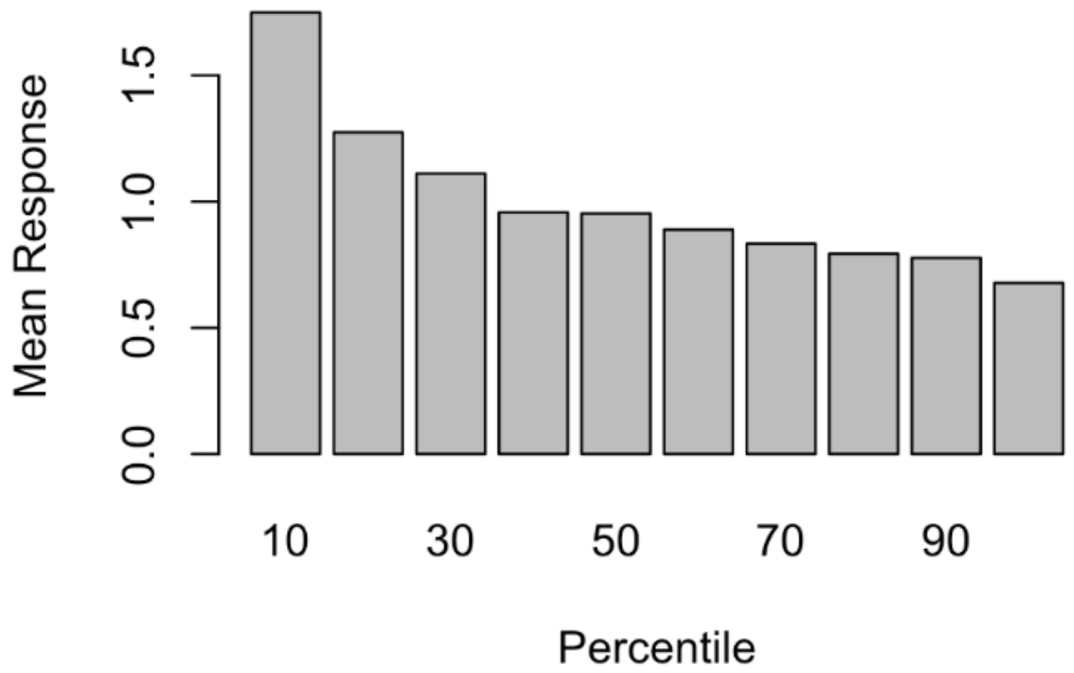
Comparing Training and Validation Performance

- The training set tell us about model fit
- The validation set (called “prediction errors”) measure the model’s ability to predict new data (predictive performance)
- We expect training errors to be smaller than the validation errors (because the model was fitted using the training set)
- The more complex the model, the greater the likelihood that it will overfit the training data (indicated by a greater difference between the training and validation errors).

Lift Chart

- The lift chart is a measure for evaluating the **performance of the classification model**.
  - It is a table that is calculated and displayed with information such as response detection rate, response rate, lift, etc. for each grade randomly divided to indicate how well the prediction was made for the classified observations.
- You can check **how much profit you can make when using the model as a percentage** compared to what you can randomly identify the object of interest to.


Lift Chart
Cumulative Price
0
2000000
0
100
200
300
400
cases


Decile-wise lift chart
Mean Response
0.0
0.5
1.0
1.5
10
30
50
70
90
Percentile

It can be useful in the following scenario:

- choosing the top 10% of the cars that gave the highest predicted sales,
    - we would gain 1.7 times the amount of revenue,
    - compared to choosing 10% of the cars at random.
- This can be seen from the decile chart
    - This number can also be computed from the lift chart by comparing the sales for 40 random cars
        - the value of the baseline curve at x = 40, which is 486,871 (= the sum of the actual sales for the 400 validation set cars divided by 10)
        - the actual sales of the 40 cars that have the highest predicted values (the value of the lift curve at x = 40), $835,883.
        - The ratio between these numbers is 1.7.

5.3 Judging Classifier Performance

- Before we study these various algorithms in detail and face decisions on how to set these options, we need to know how we will measure success.
- A natural criterion for judging the performance of a classifier
  - Probability of making a misclassification error.

Error and Error Rate

- **Error**: classifying a record as belonging to one class when it belongs to another class
- **Error rate**: percent of misclassified records out of the total records in the validation data

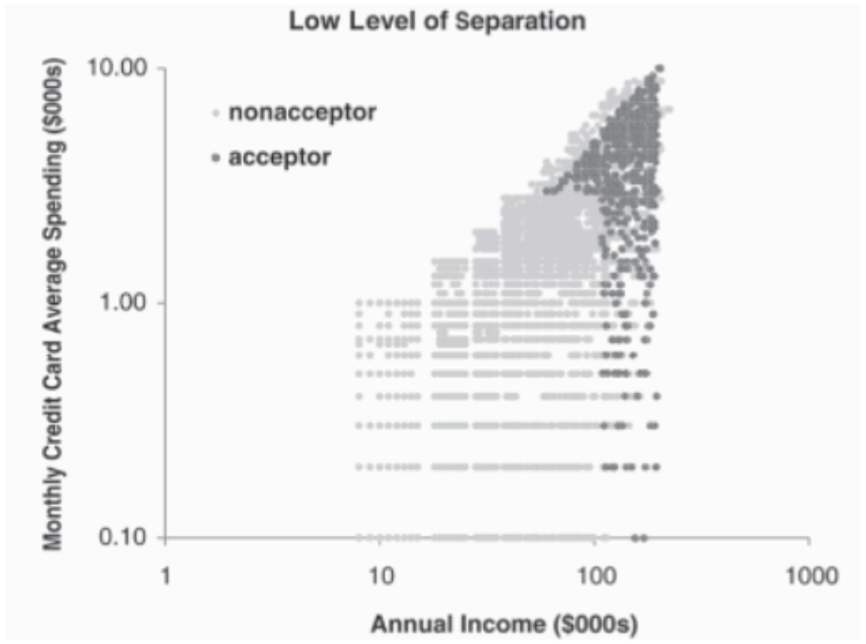
Benchmark: The Naive Rule

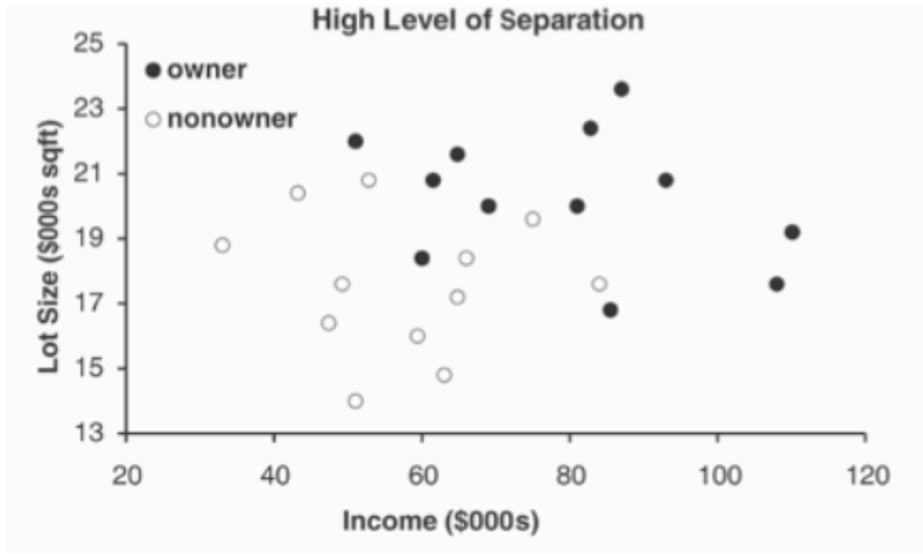
Naïve rule: classify all records as belonging to the most prevalent class

- Often used as benchmark: we hope to do better than that
- Exception: when goal is to identify high-value but rare outcomes, we may do well by doing worse than the naïve rule (see “lift” – later)

Class Separation

- **High separation of records**: using predictor variables attains low error
- **Low separation of records**: using predictor variables does not improve much on naïve rule





The Confusion (Classification) Matrix

- Most accuracy measures are derived from the confusion matrix (classification matrix) - This matrix summarizes the correct and incorrect classifications that a classifier produced for a certain dataset.
- Rows and columns of the confusion matrix correspond to the predicted and true (actual) classes, respectively.
- The confusion matrix gives estimates of the true classification and misclas sification rates.

		Actual Class	
		0	1
Predicted Class	0	2689	85
	1	25	201

Using the Validation Data

- Using the confusion matrix that is computed from the validation data.

    1. Partitioning the data into training and validation sets by random selection of records.
    2. We then construct a classifier using the training data, and then apply it to the validation data.
    3. This will yield the predicted classifications for records in the validation set
    4. Summarizing these classifications in a confusion matrix.

- Although we can summarize our results in a confusion matrix for training data as well, the resulting confusion matrix is not useful for getting an honest estimate of the misclassification rate for new data due to the danger of overfitting.

- In addition to examining the validation data confusion matrix to assess the classification performance on new data, we compare the training data confusion matrix to the validation data confusion matrix, **in order to detect overfitting**

- A large discrepancy in training and validation performance might be indicative of over-fitting.

Accuracy Measures

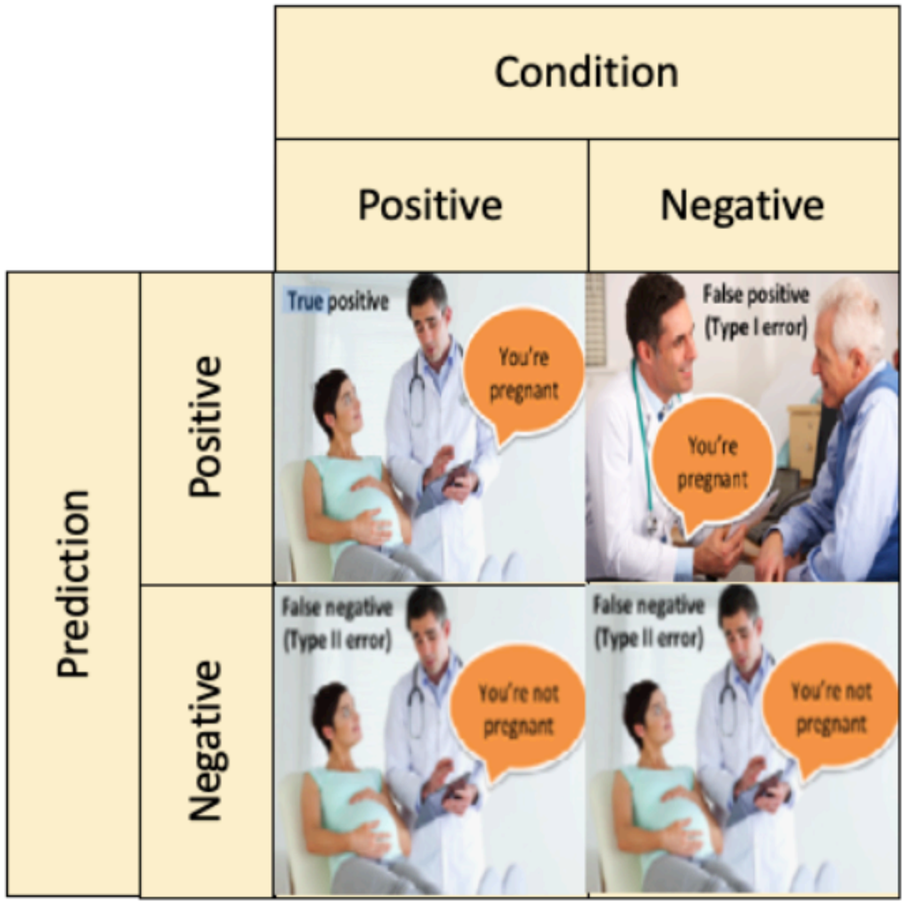
		Condition	
		Positive	Negative
Prediction	Positive	True Positive	False Positive (Type I Error)
	Negative	False Negative (Type II Error)	True Negative

		Condition	
		Positive	Negative
Prediction	Positive		False Positive (Type I Error)
	Negative	False Negative (Type II Error)	True Negative

		Condition	
		Positive	Negative
Prediction	Positive	True Positive	False Positive (Type I Error)
	Negative	False Negative (Type II Error)	

		Condition	
		Positive	Negative
Prediction	Positive	True Positive	
	Negative	False Negative (Type II Error)	True Negative

		Condition	
		Positive	Negative
Prediction	Positive	True Positive	False Positive (Type I Error)
	Negative		True Negative


Condition
Positive
Negative
Prediction
Positive
Negative
True positive
You're pregnant
False positive (Type I error)
You're pregnant
False negative (Type II error)
You're not pregnant
False negative (Type II error)
You're not pregnant

		Condition		
		Positive	Negative	
Prediction	Positive	True Positive (TP)	False Positive (FP)	Positive Predication
	Negative	False Negative (FN)	True Negative (TN)	Negative Prediction
		Sensitivity	Specificity	

		Condition	
		Positive	Negative
Prediction	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)
		Sensitivity TP / TP + FN	Specificity TN / FP + TN

Positive Predication	TP / TP + FP
Negative Prediction	TP / FN + TN

$$\text{Accuracy} = \frac{\text{True Negative (TN)} + \text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Positive (FP)} + \text{False Negative (FN)} + \text{True Negative (TN)}}$$

$$\text{Error Rate} = \frac{\text{False Positive (FP)} + \text{False Negative (FN)}}{\text{True Positive (TP)} + \text{False Positive (FP)} + \text{False Negative (FN)} + \text{True Negative (TN)}}$$

The sensitivity

The sensitivity (also termed recall) of a classifier is its ability to detect the important class members correctly. This is measured by n1,1/(n1,1 + n1,2), the percentage of C1 members classified correctly.

The specificity

The specificity of a classifier is its ability to rule out C2 members correctly. This is measured by n2,2/(n2,1 + n2,2), the percentage of C2 members classified correctly.

		Condition		
		Positive	Negative	
Prediction	Positive	True Positive 10	False Positive 90	Positive Predication 10/10+90 10%
	Negative	False Negative 5	True Negative 895	Negative Prediction 895/5+895 99.4%
		Sensitivity 10/10+5 67%	Specificity 895/90+895 90.9%	

Accuracy = 90.5%

$$\frac{\text{True Negative (TN) 895} + \text{True Positive (TP) 10}}{\text{True Positive (TP) 10} + \text{False Positive (FP) 90} + \text{False Negative (FN) 5} + \text{True Negative (TN) 895}}$$

Error Rate = 9.5%

$$\frac{\text{False Positive (FP) 90} + \text{False Negative (FN) 5}}{\text{True Positive (TP) 10} + \text{False Positive (FP) 90} + \text{False Negative (FN) 5} + \text{True Negative (TN) 895}}$$

What is good model?

Good model: high ratio in accuracy, sensitivity, specificity

Practice

You have created a device that predicts whether or not lovers will have an affair. With this device, 1000 students from Wenzhou-Kean University tested the device. A month later, 1,000 students who actually participated in the experiment were asked if their lover had cheated on them. Of the 700 people who predicted that the device you developed would cheat on your partner, 500 actually had an affair with your partner. And out of the remaining 300 who predicted that their partner would not cheat, only 100 responded that their lover did not cheat, and the rest said that their partner had cheated on them.

Evaluate the accuracy, error rate, and performance of this device.

已经创建了一个可以预测恋人是否会有外遇的设备。 温州肯恩大学的 1000 名学生使用该设备对其进行了测试。 一个月后，1000 名实际参与实验的学生被问及他们的爱人是否背叛了他们。 在预测您开发的设备会欺骗您的伴侣的 700 人中，有 500 人实际上与您的伴侣有染。 而在剩下的 300 名预测他们的伴侣不会出轨的人中，只有 100 人回答说他们的爱人没有出轨，其余的人说他们的伴侣出轨了。

评估此设备的准确性、错误率和性能。

Prediction		Condition: Positive	Condition: Negative	
	Positive	True Positive 500	False Positive 200	Positive Predication 500/500+200 71%
	Negative	False Negative 100	True Negative 200	Negative Prediction 100/100+200 33%
		Sensitivity 500/500+100 83%	Specificity 200/200+200 50%	

Accuracy = 70%

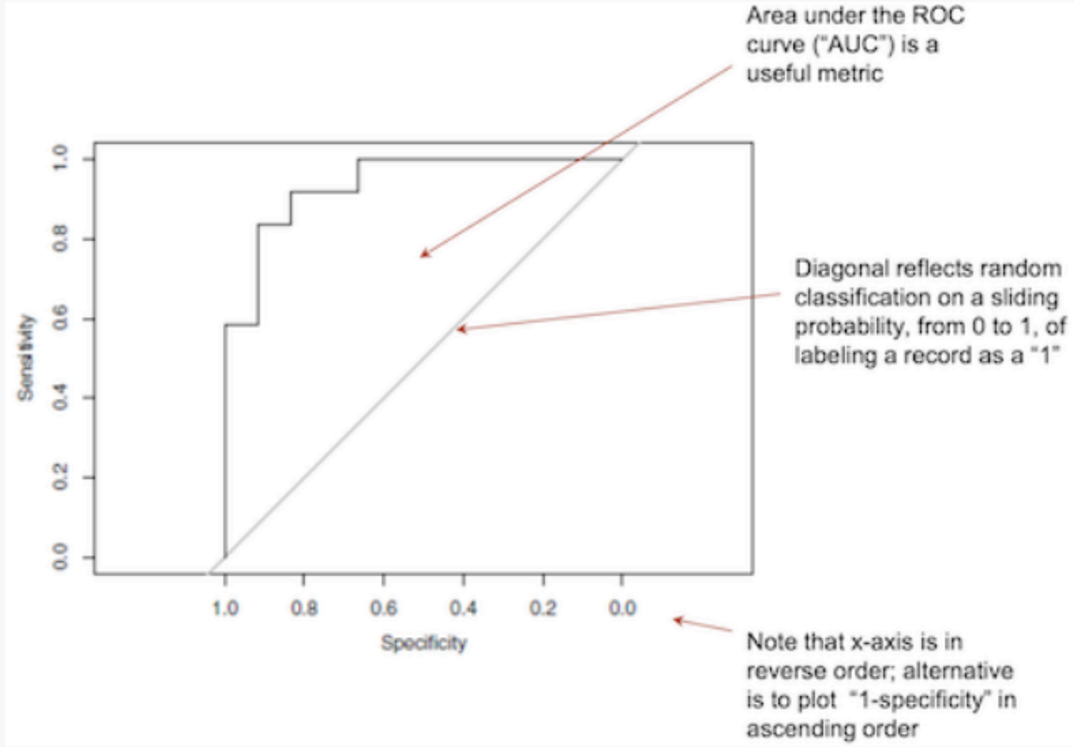
$$\frac{\text{True Negative (TN) 200} + \text{True Positive (TP) 500}}{\text{True Positive (TP) 500} + \text{False Positive (FP) 200} + \text{False Negative (FN) 100} + \text{True Negative (TN) 200}}$$

Error Rate = 30%

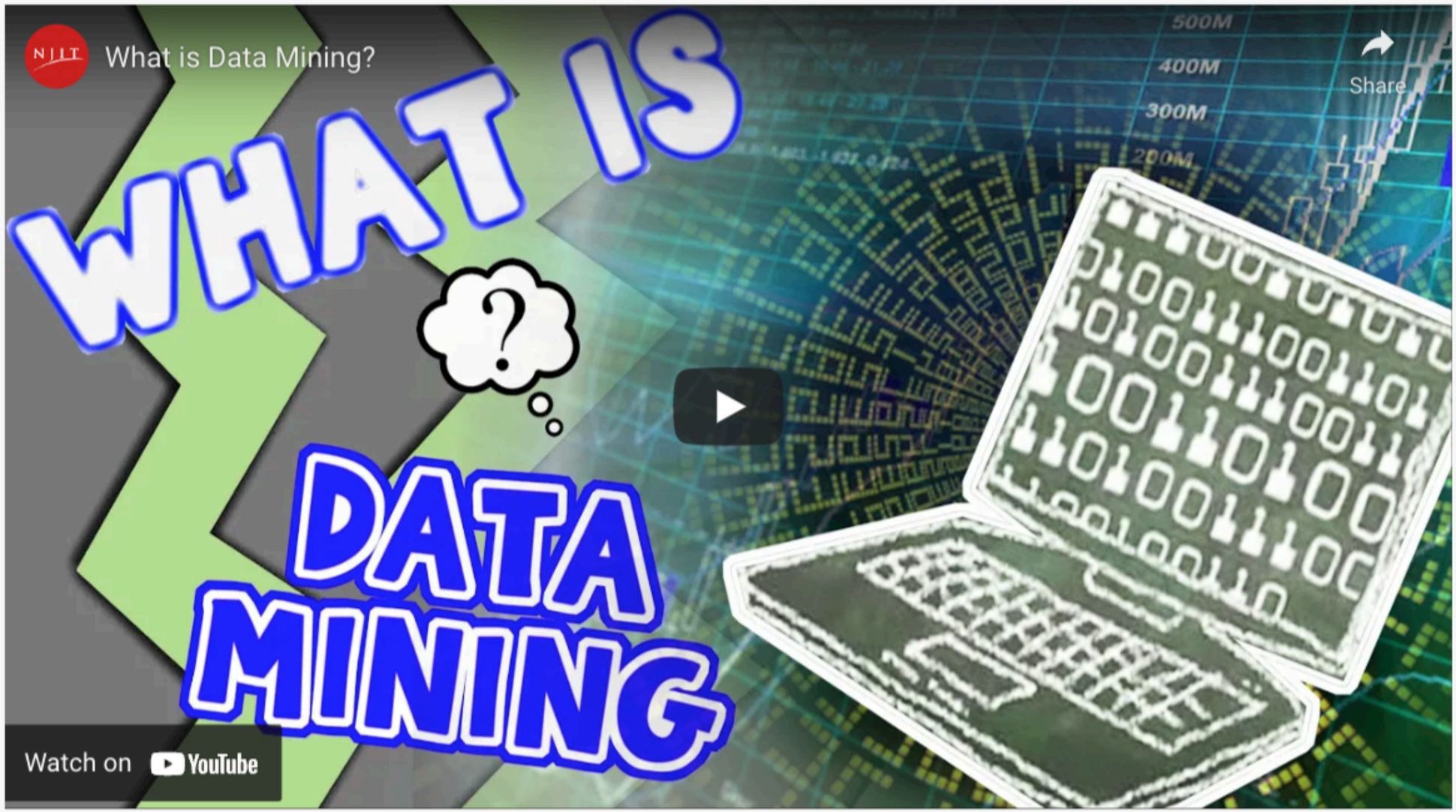
$$\frac{\text{False Positive (FP) 200} + \text{False Negative (FN) 100}}{\text{True Positive (TP) 500} + \text{False Positive (FP) 200} + \text{False Negative (FN) 100} + \text{True Negative (TN) 200}}$$

ROC Curve (Receiver Operating Characteristic)

- The ROC curve plots the pairs (sensitivity, specificity) as the cut-off value descends from 1 to 0.
- Better performance is reflected by curves that are closer to the top-left corner.
- The comparison curve is the diagonal, which reflects the performance of the naive rule
- Area under the curve (AUC) which ranges from 1 (perfect discrimination between classes) to 0.5 (no better than the naive rule).



Chapter Video



39 / 41 00:59

References

Shmueli, G., P. C. Bruce, P. Gedeck, and N. R. Patel (2019). Data mining for business analytics: concepts, techniques and applications in Python. John Wiley & Sons.

Shmueli, G., P. C. Bruce, I. Yahav, N. R. Patel, and K. C. Lichtendahl Jr (2017). Data mining for business analytics: concepts, techniques, and applications in R. John Wiley & Sons.

Assignment

Class Plan & Final Exam

- Class plan
  - May 10 - 13: Chapter6 - Regression
  - May 16 - 20: Chapter20 - Text Mining
- Range: Chapter 5, 6, 20
- May 24 (MKT)
  - (MKT Grouop1), 5:30 - 6:10
  - (MKT Grouop2), 6:15 - 6:45
- May 25 (MGS)
  - (MGS Grouop1), 2:30 - 3:10
  - (MGS Grouop2), 3:15 - 4:45
- 1 learning sheet (2 page, front and back, hand-writing only)
- No class on 26, 27, 31 (Final week)

Project Project (Project Proposal)

- Due date: May 27 (Friday 23:55pm, Beijing Time)
- To do:
  - Individual Assignment
  - 5 page (up to 7 page, no more than 7 page) project plan
  - Double space, 11 point, Times new roman
  - No reference page
  - Project proposal should includes your project plan guide by data mining process (chapter2)
  - PDF only (both R markdown generated or MS WORD)
  - Including your name and student ID and title
  - Submit to LMS (Submit button will be disappear 5 mins before midnight)