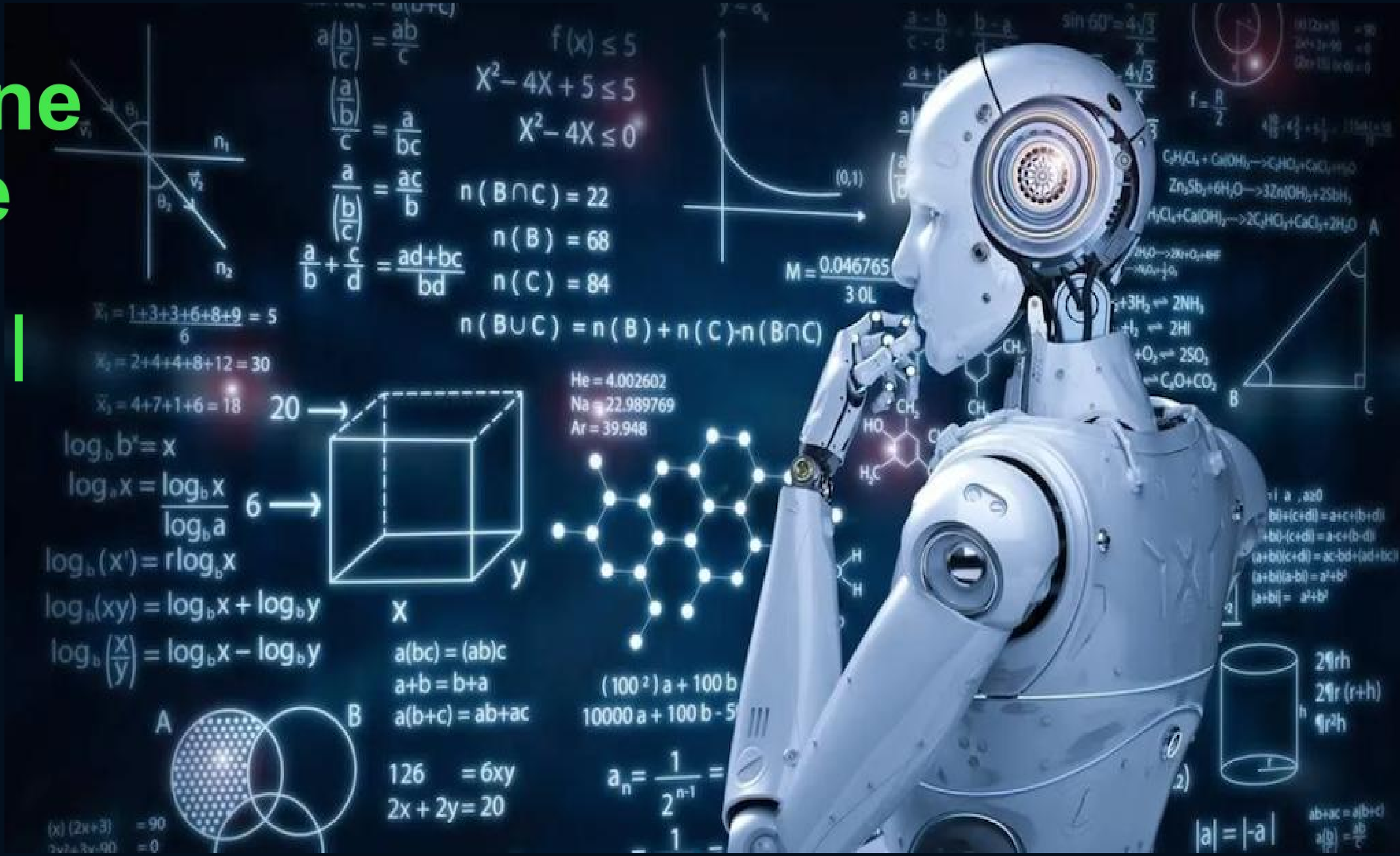
Principles of Machine Learning in Finance

4. Supervised Learning | Classification | Naive Bayes



Learning Outcomes

- Classification Models and Naive Bayes
- Conditional Probability
- Posterior Probability
- Naive Bayes in Python
- Data Scaling
- Evaluation Metrics for Naive Bayes
- **Coding Activity 4**: Supervised ML. Classification. Naive Bayes || [Naive Bayes Model with Python]

Naive Bayes

Naive Bayes is a supervised classification technique that is based on Bayes' theorem with an assumption of independence among predictors.

Posterior probability is a probability of an event occuring after taking into consideration new information.

Conditional Probability

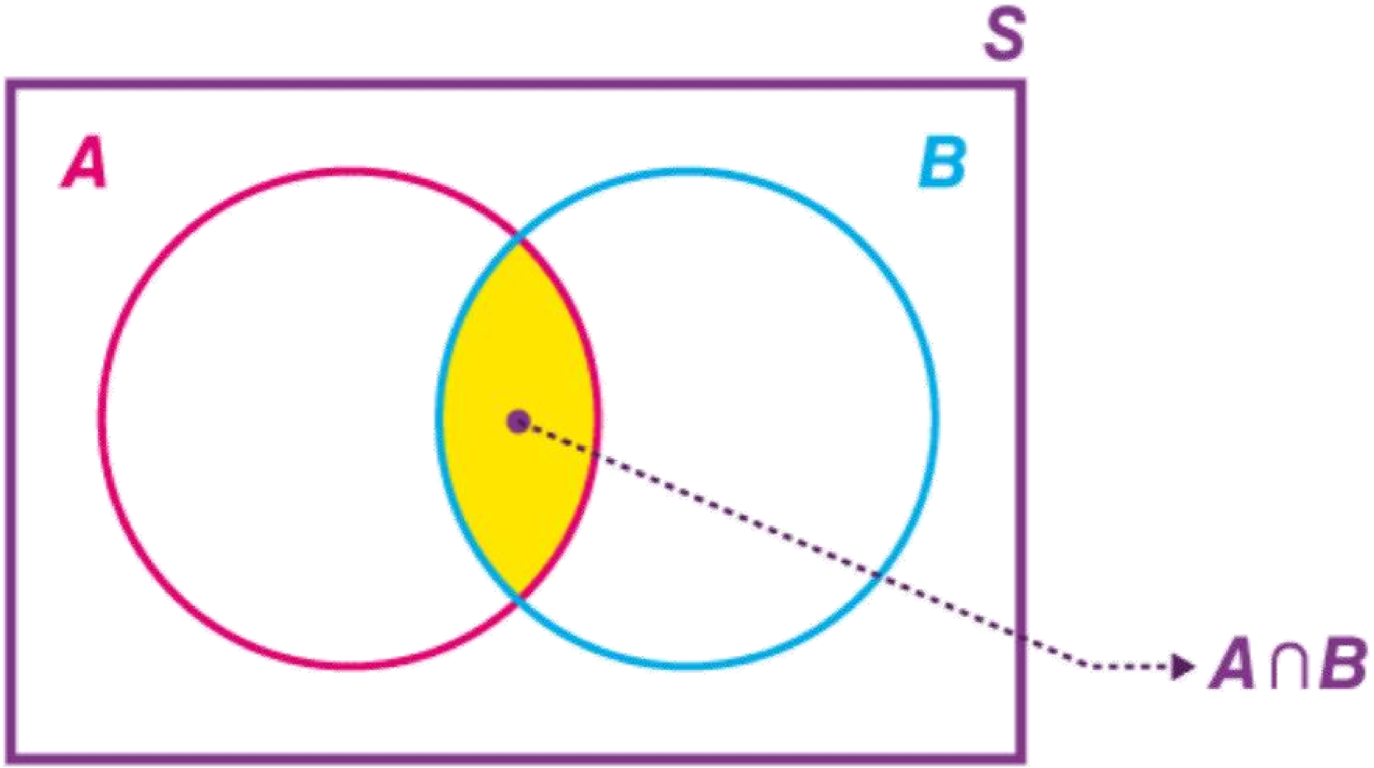
The conditional probability of an event A is the probability that the event will occur given the knowledge that an event B has already occurred.

a). In case where events A and B are **independent** (where event B has no effect on the probability of event A), the conditional probability of event A given B is $P(A \mid B) = P(A)$;

b). If event A and B are **not independent,** the conditional probability $P(A \mid B)$ equals

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, where\ P(B) > 0$$

Conditional Probability: Venn Diagram



$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Example 1. Conditional Probability

Question 1: The percentage of adults who are male and traders is 2,25%. What is the probability of being a trader, given being a male?

Notation: A = trader and B = male

Example 1. Conditional Probability: Solution

Question 1: The percentage of adults who are male and traders is 2,25%. What is the probability of being a trader, given being a male?

Notation: A = trader and B = male

$$P(A \cap B) = 0.0225$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0,0225}{0,5} = 0,045$$

Example 2. Conditional Probability

Question 2: What is the probability of two assets are stocks given at least one is a stock?

Example 2. Conditional Probability: Solution

Question 2: What is the probability of two assets are stocks given at least one is a stock?

$$P(2S \mid 1S) = \frac{P(2S \cap 1S)}{P(1S)}$$

$$Posibilities = \{SS, SN, NS, NN\}$$

$$P(2S \mid 1S) = \frac{1/4}{3/4} = \frac{1}{3}$$

Posterior Probability and The Bayes' Theorem

Bayes Theorem describes the probability of occurrence of an event related to any condition. Bayes' theorem is known as the formula for the probability of "causes".

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)}$$

$$P(A \mid B) = \frac{P(B \cap A) \cdot P(A)}{P(B)}, \; where \; P(B) \neq 0$$

Example 3. The Bayes' Theorem

Question 3: What is the probability of two assets are stocks given at least one is a stock?

$$P(2S \mid 1S) = \frac{P(1S \mid 2S) \bullet P(2S)}{P(1S)}$$

$$Posibilities = \{SS, SN, NS, NN\}$$

Example 3. The Bayes' Theorem: Solution

Question 3: What is the probability of two assets are stocks given at least one is a stock?

$$P(2S \mid 1S) = \frac{P(1S \mid 2S) \cdot P(2S)}{P(1S)}$$

$$Posibilities = \{SS, SN, NS, NN\}$$

$$P(2S \mid 1S) = \frac{1 \cdot 1/4}{3/4} = \frac{1}{3}$$

Example 4. Posterior Probability and Bayes’ Theorem

Question 4: There are two investment portfolios with six stocks each. Only two kinds of stocks are included: Apple and Samsung. Portfolio 1 includes three stocks of each types. Portfolio 2 includes two Apple and four Samsung stocks. If you randomly pick an Apple stock, what is the probability of Apple stock being in the Portfolio 1?

Notation: A = select an Apple stock; B1 = Portfolio 1; B1 = Portfolio 2

Example 4. Posterior Probability and Bayes' Theorem: Solution (1)

1. $P(A \mid B1) = \dfrac{1}{2}$

2. $P(A \mid B2) = \dfrac{1}{3}$

3. $P(B1) = P(B2) = \dfrac{1}{2}$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Example 4. Posterior Probability and Bayes' Theorem: Solution (2)

4. $$P(A) = P(A \cap B1) + P(A \cap B2) =$$
$$= P(A|B1) \bullet P(B1) + P(A|B2) \bullet P(B2) =$$
$$= \frac{1}{2} \bullet \frac{1}{2} + \frac{1}{3} \bullet \frac{1}{2} = \frac{1}{4} + \frac{1}{6} = \frac{5}{12}$$

5. $$P(B1 \mid A) = \frac{P(A \mid B1) \bullet P(B1)}{P(A)} = \frac{\frac{1}{2} \bullet \frac{1}{2}}{\frac{5}{12}} = \frac{3}{5}$$

Posterior Probability

Likelihood of a predictor x given a class c

Class prior probability

$$P(c|x) = \frac{P(x|c)\,P(c)}{P(x)}$$

Posterior probability

Predictor prior probability

Posterior Probability

$$P(c|x) = \frac{P(x|c)\,P(c)}{P(x)}$$

$$\underbrace{P(c \mid X)}_{\text{Posterior probability}} = \underbrace{P(x_1 \mid c) * P(x_2 \mid c) * \ldots P(x_n \mid c)}_{\text{Conditional probabilities}} * \underbrace{P(c)}_{\text{Probability of a class c}}$$

Example 5. Posterior Probability

Outlook	Temp	Humidity	Windy	Play Football
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Example 5. Posterior Probability (1)

Play

Frequency Table		Play Football	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

$P(x \mid c) = P(\text{Sunny} \mid \text{Yes}) = 3/9 = 0.33$

Likelihood Table		Play Football		
		Yes	No	
Outlook	Sunny	3/9	2/5	5/14
	Overcast	4/9	0/5	4/14
	Rainy	2/9	3/5	5/14
		9/14	5/14	

$P(x) = P(\text{Sunny}) = 5/14 = 0.36$

$P(c) = P(\text{Yes}) = 9/14 = 0.64$

Posterior Probability: $P(c \mid x) = P(\textbf{Yes} \mid \text{Sunny}) = 0.33 \times 0.64 \div 0.36 = 0.60$

Example 5. Posterior Probability (2)

Don’t Play

Frequency Table		Play Football: Yes	Play Football: No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

$P(x \mid c) = P(\text{Sunny} \mid \text{Yes}) = 3/9 = 0.33$

Likelihood Table		Play Football: Yes	Play Football: No	
Outlook	Sunny	3	2	5
	Overcast	2	0	4
	Rainy	2	3	5
		9	5	

$P(x) = P(\text{Sunny}) = 5/14 = 0.36$

$P(c) = P(\text{Yes}) = 9/14 = 0.64$

Posterior Probability: $P(c \mid x) = P(\textbf{No} \mid \text{Sunny}) = 0.40 \times 0.36 \div 0.36 = 0.40$

Naive Bayes Implementaion in Scikit-learn

There are several implementations of Naive Bayes in scikit-learn, all of which are found in the **sklearn.naive_bayes** module:

- **BernoulliNB**: Used for binary/Boolean features
- **CategoricalNB**: Used for categorical features
- **ComplementNB**: Used for imbalanced datasets, often for text classification tasks
- **GaussianNB**: Used for continuous features, normally distributed features
- **MultinomialNB**: Used for multinomial (discrete) features

Key Evaluation Metrics: Accuracy and Precision

Accuracy reflects the number of correct predictions divided by the total number of predictions:

$$Accuracy = \frac{N_{correct\ predictions}}{N_{total\ predictions}} = \frac{True\ positives + True\ negtives}{N_{total\ predictions}}$$

Precision is a proportion of positive predictions that were correct to all positive predictions:

$$Precision = \frac{True\ positives}{True\ positives\ +\ False\ positives}$$

Key Evaluation Metrics: Recall

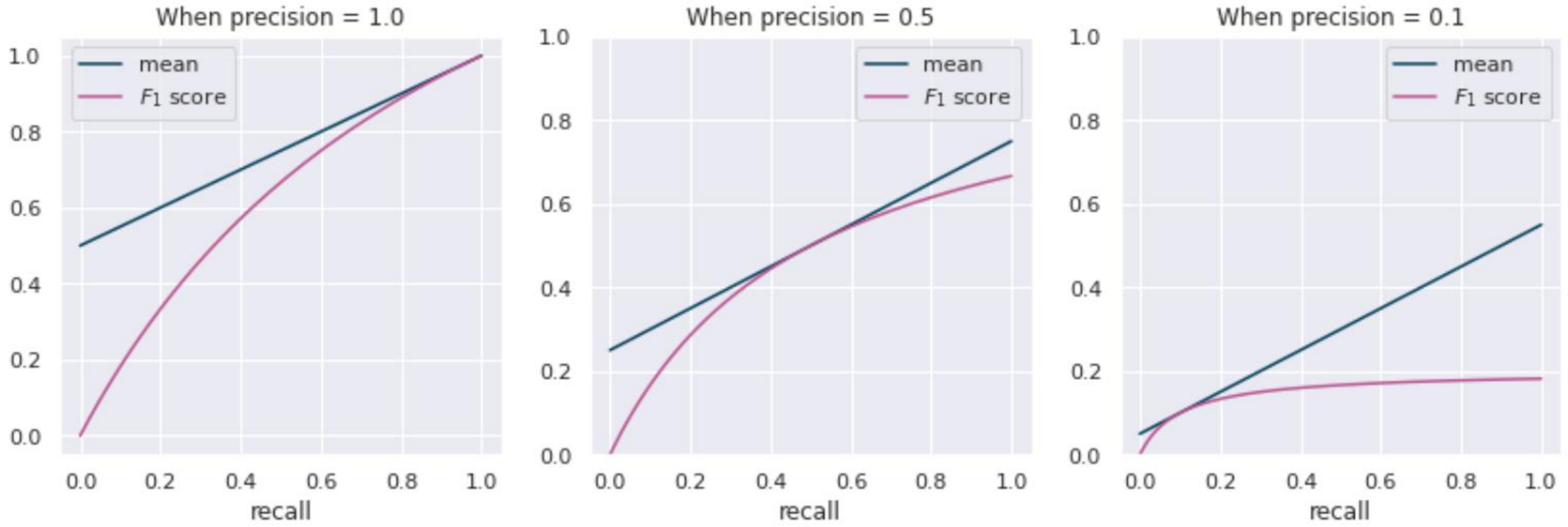
Recall is a proportion of actual positives that were identidied correctly to all actual positives

$$Recall = \frac{True\ Positives}{True\ Positives\ +\ False\ Negatives}$$

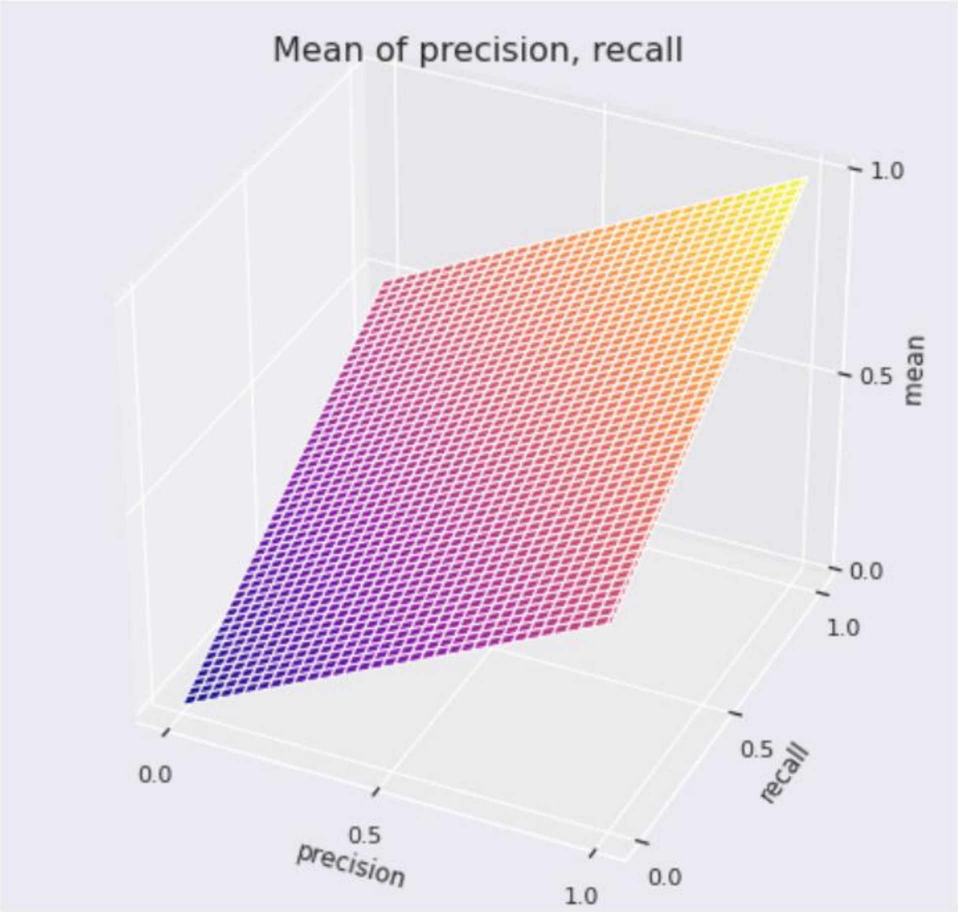
Key Evaluation Metrics: F1 score

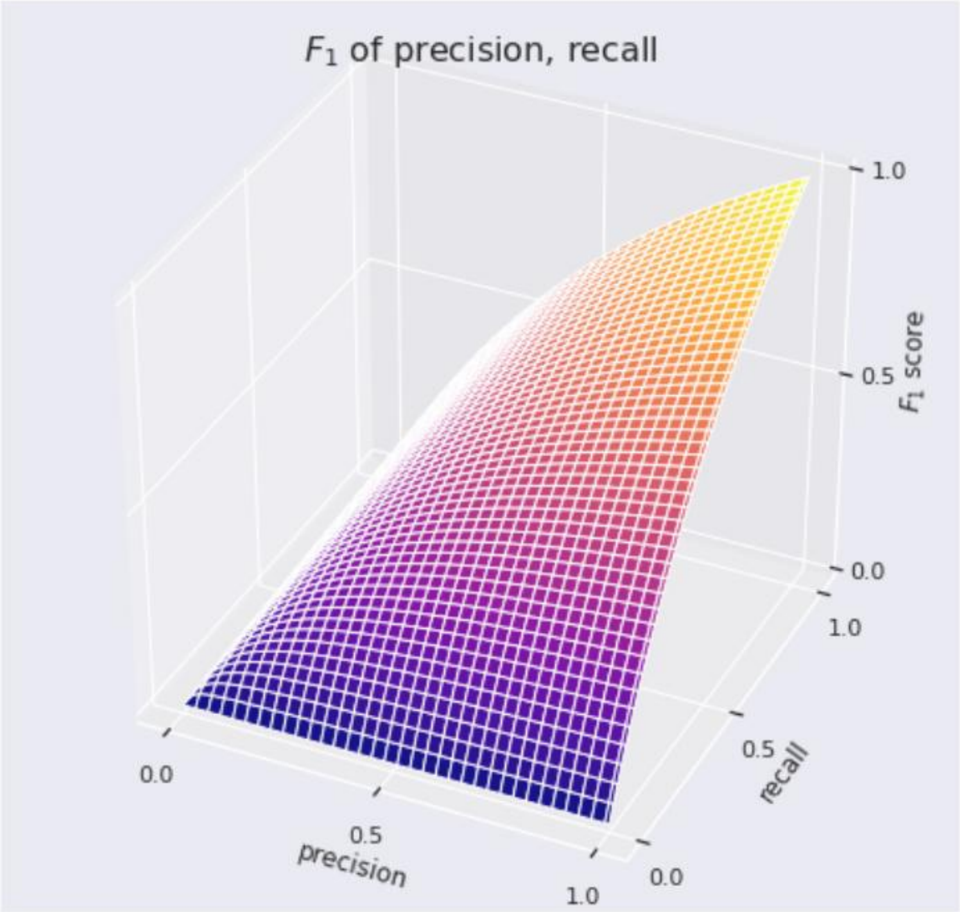
F1 score is the harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

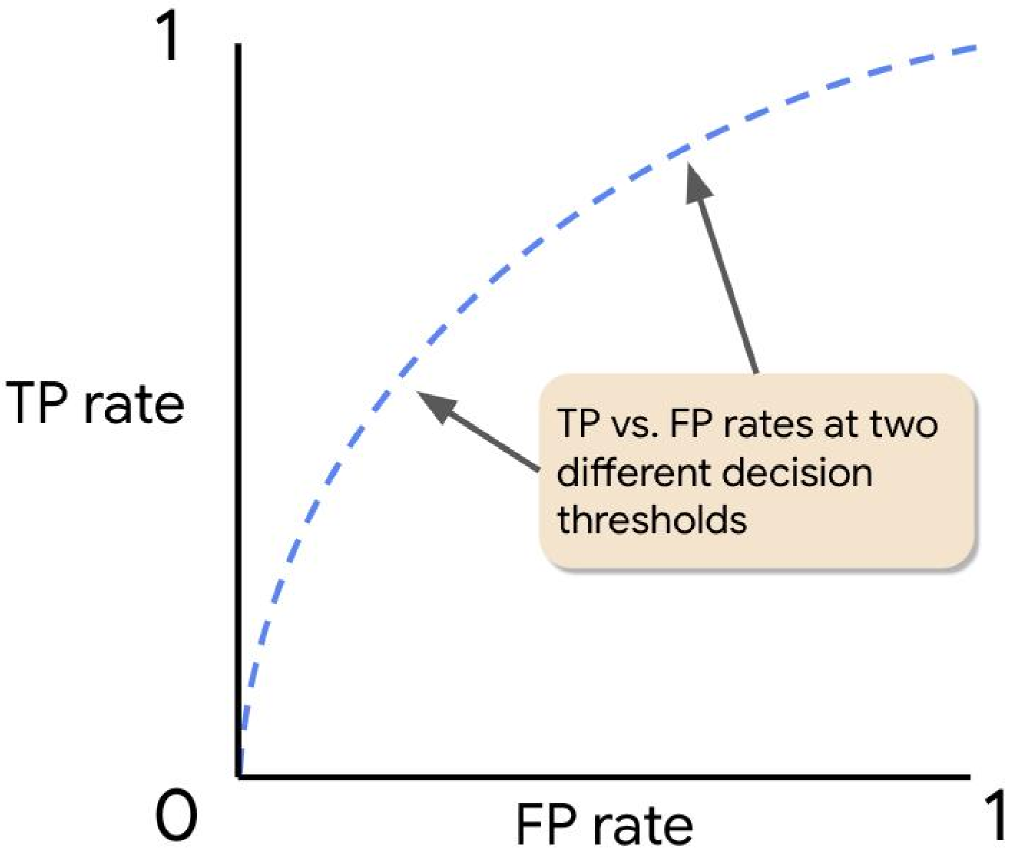


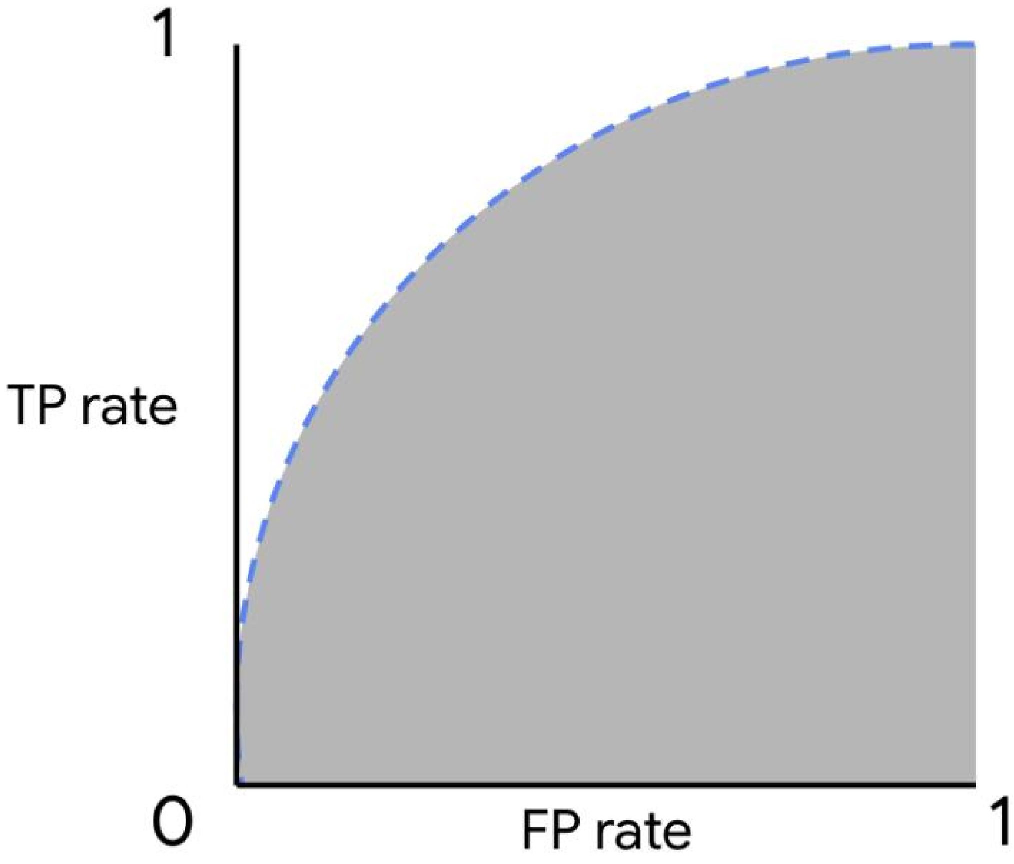
Key Evaluation Metrics: F1 score (2)





Key Evaluation Metrics: ROC Curve and AUC





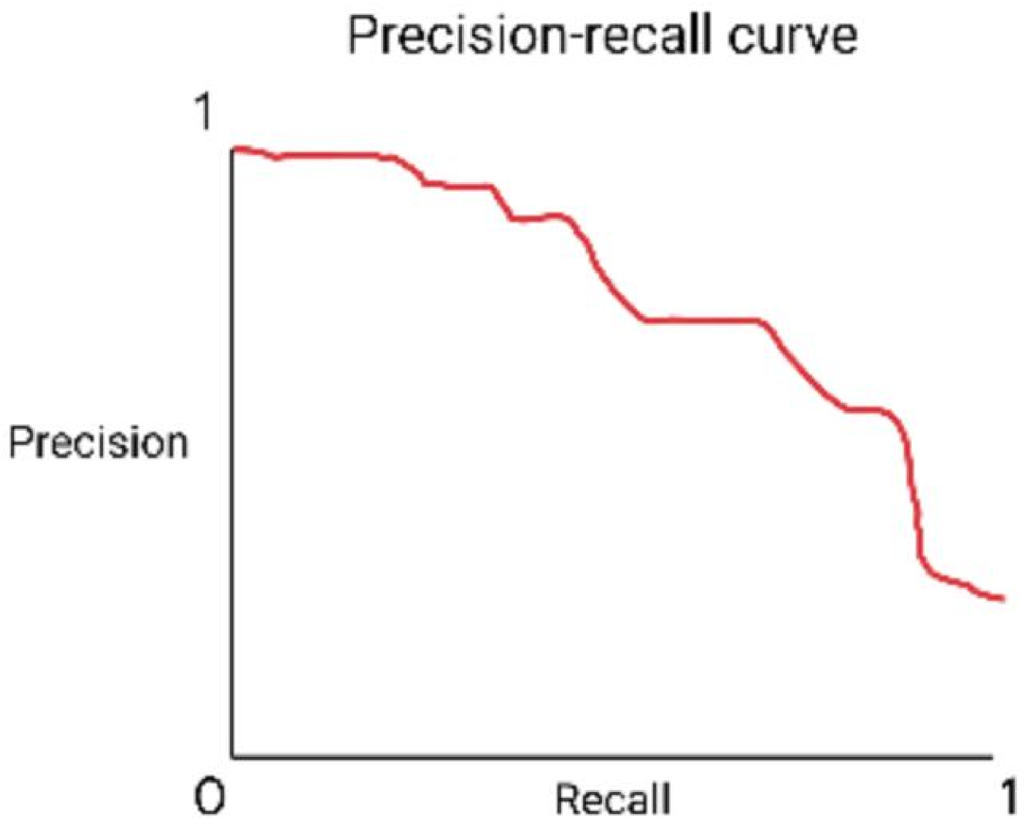
Key Evaluation Metrics: F_β score

$$F_\beta = (1 + \beta^2) \bullet \frac{Precision \bullet Recall}{(\beta^2 \bullet Precision) + Recall}$$

In F_β score $\boldsymbol{\beta}$ is a factor that represents how many times more important recall is compared to precision

Key Evaluation Metrics: Precision-Recall curves

- **Precision-recall curve** is a way to visualize the performance of a classifier at different decision thresholds (**classification thresholds**).
- In the context of binary classification, a decision threshold is a probability cutoff for differentiating the positive class from the negative class.



Coding Activity 4. Supervised ML. Classification

Lab 4. Supervised Machine Learning. Classification. Naive Bayes || [Naive Bayes Model with Python]

Steps to follow:

1. Upload the following files from the module learning room:
   - Jupiter notebook “Lab4_Naive_Bayes_with_Python.ipynb”
   - Csv-dataset file “BankModelling_FE.csv”
2. Follow along in the Jupiter notebook

Coding Activity 4. Dataset Overview

	Credit Score	Age	Tenure	Balance	NumOf Products	Has CrCard	IsActive Member	Estimated Salary	Exited	Loyalty	Geography_ Germany	Geography_ Spain
0	619	42	2	0.00	1	1	1	101348.88	1	0.047619	0	0
1	608	41	1	83807.86	1	0	1	112542.58	0	0.024390	0	1
2	502	42	8	15966.80	3	1	0	113931.57	1	0.190476	0	0
3	699	39	1	0.0	2	0	0	93826.63	0	0.025641	0	0
4	850	43	2	125510.82	1	1	1	79084.10	0	0.046512	0	1

Coding Activity 2. Scaling Techniques

- Some models require you to scale the data in order for them to operate as expected, while others don’t.
- In general, Naive Bayes does not require data scaling.

Min-max scaler

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}; \quad X_{scaled} \in [0;\ 1]$$

Q&A

Thank you!