

Day 17 – ML Theory

Model Evaluation Metrics (Classification)

Amaç

Bu çalışmanın amacı, **classification problemlerinde model performansını doğru metriklerle değerlendirmeyi** öğrenmek ve **iş problemine göre hangi metriğin neden seçilmesi gerektiğini** kavramaktır.

Gün 16'daki *Cross-Validation & Data Leakage* konusunun doğal devamıdır. Burada artık **"modelim iyi mi?"** sorusuna **doğru metrikle** cevap vermeyi öğreniyoruz.

1. Confusion Matrix Nedir?

Binary classification için temel yapı:

	Tahmin: 1	Tahmin: 0
Gerçek: 1	TP	FN
Gerçek: 0	FP	TN

- **TP (True Positive):** Gerçek 1 → Tahmin 1
- **FP (False Positive):** Gerçek 0 → Tahmin 1
- **FN (False Negative):** Gerçek 1 → Tahmin 0
- **TN (True Negative):** Gerçek 0 → Tahmin 0

🔗 Tüm classification metrikleri **bu 4 sayıdan** türetilir.

2. Accuracy (Doğruluk)

$$\text{Accuracy} = (TP + TN) / (TP + FP + FN + TN)$$

Ne Zaman İyi?

- Sınıflar **dengeli** ise
- Yanlış pozitif ve yanlış negatif **aynı derecede önemliyse**

Neden Tek Başına Yetersiz?

Imbalanced data'da **yanıltıcıdır**.

🔗 Örnek (Fraud / Churn):

- %95 müşteri churn değil

- Model herkes için “churn değil” derse:
 - Accuracy = %95
 - Model **işe yaramaz**

➡ Accuracy **tek başına karar metriği OLAMAZ.**

3. Precision (Kesinlik)

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

Ne Soruyor?

“1 dediğim tahminlerin kaç gerçekten 1?”

Ne Zaman Kritik?

- **False Positive maliyeti yüksekse**
- Yanlış alarm vermek istemiyorsak

🔗 Örnek:

- Spam filtresi
- Fraud flag'leri
- Yanlış müşteriye churn diye etiketlemek

➡ Precision yüksek → “Emin olmadıkça 1 deme”

4. Recall (Duyarlılık / Sensitivity)

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

Ne Soruyor?

“Gerçek 1’lerin ne kadarını yakalayabildim?”

Ne Zaman Kritik?

- **False Negative maliyeti yüksekse**
- Bir şeyi kaçırmak çok pahalıysa

🔗 Örnek:

- Kanser taraması
- Fraud detection
- Riskli müşteriye kaçırmak

📌 Recall yüksek → “Hiçbir 1’i kaçıрма”

5. Precision – Recall Trade-off

- Precision ↑ → Recall genelde ↓
- Recall ↑ → Precision genelde ↓

Bu trade-off:

- **Decision threshold** ile kontrol edilir
- Varsayılan threshold = **0.5**

🔗 Business kararı:

- Fraud → Recall odaklı
 - Marketing campaign → Precision odaklı
-

6. F1-Score

$$F1 = 2 * (Precision * Recall) / (Precision + Recall)$$

Ne Yapar?

- Precision ve Recall’un **denge noktası**
- İkisini de optimize etmeye çalışır

Ne Zaman Kullanılır?

- Sınıflar dengesizse
- FP ve FN **benzer derecede önemliyse**

🔗 Churn, risk, credit scoring gibi problemler için **en sık kullanılan metriktir.**

7. ROC Curve Nedir?

ROC Curve:

- X eksen: **False Positive Rate (FPR)**
- Y eksen: **True Positive Rate (Recall)**

Modelin:

- Farklı threshold’larda
 - Ne kadar iyi ayırt ettiğini gösterir
-

8. AUC (Area Under Curve)

- ROC eğrisinin altındaki alan
- 0.5 → rastgele tahmin
- 1.0 → kusursuz ayırım

Ne Anlatır?

“Model, rastgele bir pozitif ve negatif örneği ayırt edebiliyor mu?”

Avantajları

- Threshold bağımsız
- Sınıf dengesizliğinden daha az etkilenir

✂ Ama:

- Business maliyetlerini **doğrudan yansıtmaz**

9. Hangi Metrik Ne Zaman?

Senaryo	Öncelikli Metrik
Dengeli sınıflar	Accuracy
Yanlış alarm pahalı	Precision
Kaçırmak pahalı	Recall
Genel denge	F1-score
Model karşılaştırma	ROC-AUC

➡ “En iyi model” diye tek bir metrik yoktur.

10. Gerçek Projelerde Nasıl Anlatılır?

Mülakatta güçlü cevap:

“Accuracy yerine F1 ve ROC-AUC kullandım çünkü dataset imbalanced’tı. Business tarafında FN maliyeti yüksek olduğu için recall’ı ayrıca takip ettim. Threshold’u default 0.5 yerine validation set üzerinden optimize ettim.”

11. Gün 17 – Kısa Özet

- Accuracy tek başına güvenilirmez
- Precision & Recall business maliyetiyle ilgilidir
- F1 dengeli bir metriktir
- ROC-AUC modelin ayırt ediciliğini ölçer
- Doğru metrik = doğru karar

✂ İyi model = yüksek skor değil, doğru metrikle ölçülmüş model

