

Day 19 – ML Theory

Model Calibration (Probability Calibration)

1. Calibration Nedir?

Classification modelleri çoğu zaman yalnızca sınıf etiketi (0 / 1) değil, aynı zamanda **olasılık (probability)** üretir.

Örnek: $P(y = 1) = 0.80$

Calibration şu soruyu sorar:

“Model 0.80 dediğinde, bu gerçekten %80 ihtimalle mi doğru?”

Eğer cevap evet ise → model **iyi kalibre edilmiştir**.

2. Accuracy Yüksek Ama Calibration Kötü Olabilir

Bir model:

- Accuracy / F1 / AUC açısından iyi olabilir
- Ama ürettiği olasılıklar **gerçekçi olmayabilir**

Örnek:

- Model sürekli 0.99 ve 0.01 üretiyor
- Ama gerçekte bu kadar emin olunmaması gerekir

➡ Bu durumda model:

- Karar vermede
 - Risk hesaplamada
 - Threshold seçiminde yanıltıcı olur.
-

3. Calibration Neden Önemlidir?

Calibration özellikle şu durumlarda kritiktir:

- Credit risk (temerrüt olasılığı)
- Churn probability
- Fraud probability
- Medical risk scoring

🔗 Business genelde şunu ister:

“Bu müşteri % kaç ihtimalle churn eder?”

Yanlış kalibre edilmiş bir model:

- Yanlış risk algısı yaratır
 - Yanlış threshold kararlarına yol açar
-

4. Calibration vs Discrimination

İki kavram **kariştirilmamalıdır**:

- **Discrimination**
 - Pozitif ve negatifleri ayırabilme gücü
 - ROC-AUC ile ölçülür
- **Calibration**
 - Olasılıkların doğruluğu
 - Reliability ile ilgilidir

🔗 Bir model:

- Yüksek AUC'ye sahip olabilir
 - Ama kötü kalibre edilmiş olabilir
-

5. Calibration Nasıl Ölçülür?

5.1 Calibration Curve (Reliability Diagram)

Adımlar:

1. Tahmin edilen olasılıkları bin'lere ayır
2. Her bin için:
 - Ortalama tahmin
 - Gerçek pozitif oranı
3. İdeal durumda:
 - Eğri, $y = x$ çizgisine yakın olur

Yorum:

- Eğri yukarıdaysa → model **under-confident**
 - Eğri aşağıdaysa → model **over-confident**
-

5.2 Brier Score

Brier Score:

Özellikleri:

- Daha düşük = daha iyi
- Hem accuracy hem calibration bilgisini içerir
- Probability-based bir metriktir

6. Modeller Neden Kötü Kalibre Olur?

Yaygın sebepler:

- Aşırı güçlü / kompleks modeller
 - Random Forest
 - Gradient Boosting
 - XGBoost
- Imbalanced dataset
- Overfitting
- Küçük validation set

🔗 Logistic Regression genelde:

- Daha **iyi kalibre**
- Ama ayırt ediciliği (AUC) daha düşük olabilir

7. Calibration Yöntemleri

7.1 Platt Scaling

- Logistic regression ile yeniden ölçekleme
- Binary classification için yaygın
- Validation set üzerinde uygulanır

7.2 Isotonic Regression

- Non-parametrik
- Daha esnek
- Küçük veri setlerinde overfitting riski vardır

🔗 Calibration **her zaman validation set üzerinde** yapılmalıdır.

8. Calibration ve Threshold İlişkisi

- Threshold kararı, olasılıklara dayanır
- Olasılıklar yanlışsa:
 - En iyi threshold bile yanlış sonuç verir

➡ Bu yüzden:

1. Önce calibration
2. Sonra threshold optimization doğru akıştır.

9. Gerçek Projelerde Nasıl Anlatılır?

Mülakatta güçlü bir ifade:

“Modelim ROC-AUC açısından güçlüydü ancak probability'lerin aşırı confident olduğunu gördüm. Calibration curve ve Brier Score ile kontrol ettim. Isotonic regression ile kalibrasyon uyguladıktan sonra threshold kararlarını yeniden optimize ettim.”

10. Day 19 – Kısa Özet

- Calibration = olasılıkların doğruluğu
- AUC yüksek olması yeterli değildir
- Risk ve karar problemlerinde kritiktir
- Calibration → Threshold → Business kararı sıralaması izlenmelidir