Woman in Data Science

Lecture 4

09.12.2025

Feature Engineering nedir?

Kısaca: Ham veriden yeni, daha anlamlı özellikler (değişkenler) üretme işlemidir.

Bir makine öğrenmesi modelinin başarısını en çok etkileyen şey genelde model değil, kullandığın özelliklerin kalitesidir.

📌 Feature Engineering'in Amacı

- Modelin veriyi daha iyi anlamasını sağlamak
- Gizli ilişkileri ortaya çıkarmak
- Predictive gücü düşük ham veriyi daha anlamlı hale getirmek
- Eksik bilgi içeren kolonları zenginleştirmek
- Karmaşık veriyi sadeleştirmek veya dönüştürmek

Aile Büyüklüğü oluşturmak (FamilySize)

SibSp + Parch + 1

→ Aile ile seyahat edenlerin hayatta kalma şansı daha yüksek

İsimden Title çıkarmak (Mr, Miss, Mrs...)

Name kolonu uzun ve gereksiz, ama içinden unvanı çektiğinde çok anlamlı olur.

Cabin'in ilk harfi

Cabin → A, B, C, D...

Yolcu güvertesi aynı olanların hayatta kalma oranı farklı olabilir.

Bilet grupları (aynı Ticket kullanan yolcular)

Aynı bilet numarası = birlikte seyahat

→ hayatta kalma olasılığı grup halinde değişebilir

Fare binning (ücreti kategoriye dökmek)

Fare çok değişken → 4 kategoriye ayırmak sinyali güçlendirir.

Age binning (yaş grupları)

0–12, 13–18, 18–35…

→ Çocukların daha iyi korunmuş olması gibi örüntüler ortaya çıkar.

Kategorik değişkenleri encode etmek

Sex → 0/1

Embarked → One-Hot Encode

📌 Özet

Feature engineering, bir modelin performansını artırmak için veriyi yeniden tasarlamak demektir.
Bir ML projesinin belki de %60'ı Feature Engineering aşamasıdır.

FEATURE SELECTION

(Filter – Wrapper – Embedded Yöntemleri)

⭐ Feature Selection Nedir?

Feature Selection, bir veri setindeki tüm özellikler arasından model için en bilgilendirici olanları seçme sürecidir.

Amaç:

- **Gereksiz (irrelevant) değişkenleri elemek**
- **Gürültüyü azaltmak**
- **Overfitting'i azaltmak**
- **Modeli hızlandırmak**
- **Performansı artırmak**

Feature Engineering = Yeni özellik üretme

Feature Selection = Hangilerinin kullanılacağına karar verme

🧭 Feature Selection Yöntemleri 3 Ana Gruba Ayrılır

1) Filter Methods (Filtre Yöntemleri)

📌 **Model bağımsızdır.**
Özelliklerin hedefle olan istatistiksel ilişkisine bakıp sıralama yapar.

Çalışma mantığı:

- **Her bir feature, hedef değişken ile tek başına değerlendirilir.**
- **İstatistiksel bir skor hesaplanır.**
- **Skoru düşük olan özellikler atılır.**

En yaygın yöntemler:

- **Correlation (Pearson, Spearman) → Sürekli değişkenler için**
- **Chi-square → Kategorik değişkenler için**
- **ANOVA F-test → Sürekli → kategorik hedef**
- **Mutual Information (MI)**
- **Variance Threshold**
- **→ Çok düşük varyansa sahip özellikleri atar**

🧭 Feature Selection Yöntemleri 3 Ana Gruba Ayrılır

⚓ Avantajlar:

- Çok hızlıdır
- Büyük veri setlerinde iyi çalışır
- Model bağımsız olduğu için her modelde kullanılabilir

⚠️ Dezavantajlar:

- Feature'ların birbirleriyle olan etkileşimlerini hesaba katmaz
- Tek tek bakar → "kollektif etkileri" göremez

Titanic Örneği:

- Sex → Survived ile çok güçlü korelasyona sahiptir
- Fare → orta düzeyde
- Ticket gibi bazı kolonların korelasyonu düşüktür → elenebilir

🧭 Feature Selection Yöntemleri 3 Ana Gruba Ayrılır

2) Wrapper Methods (Sarma Yöntemleri)

📌 **Model bağımlıdır.**

Bir model oluşturulur ve farklı feature subset'leri denenir.

Amaç: en yüksek performansı veren feature set'i bulmak.

Mantık:

- **"Bu özellik setiyle model daha iyi mi çalışıyor?" diye test eder.**

En popüler wrapper yöntemleri:

- 🔹 **Forward Selection**
  - **Boş liste ile başlar**
  - **Her adımda performansı artıran en iyi feature eklenir**
- 🔹 **Backward Elimination**
  - **Tüm özelliklerle başlar**
  - **En az katkı veren özellikler tek tek çıkarılır**
- 🔹 **Recursive Feature Elimination (RFE)**
  - **Model her iterasyonda eğitilir**
  - **En düşük önem skoruna sahip feature silinir**
  - **Tekrarlanır**

🧭 Feature Selection Yöntemleri 3 Ana Gruba Ayrılır

⚓ Avantajlar:

- Feature'lar arasındaki etkileşimi dikkate alır
- Genellikle en yüksek performansı üretir

⚠️ Dezavantajlar:

- Çok yavaştır
- Büyük veri setlerinde maliyetli
- Overfitting riski vardır

Titanic Örneği:

RFE + Logistic Regression

→ Genelde Sex, Pclass, Fare, Title gibi özellikleri seçer.

Cabin, Ticket çoğunlukla elenir.

🧭 Feature Selection Yöntemleri 3 Ana Gruba Ayrılır

3) Embedded Methods (Gömülü Yöntemler)

📌 **Model eğitimi sırasında feature importance otomatik çıkarılır.**

Yani:

- **Feature selection = modelin kendi iç mekanizmasının parçasıdır.**

En popüler gömülü yöntemler:

- 🔷 **Lasso Regression (L1 Regularization)**
  - **Bazı katsayıları tam sıfıra indirir**
  - **→ böylece feature'ı otomatik olarak silmiş olur.**
- 🔷 **Ridge Regression (L2 Regularization)**
  - **Katsayıları küçültür ama sıfırlamaz**
  - **→ zayıf sinyalli özellikleri azaltır**
- 🔷 **Decision Trees / Random Forest Importance**
  - **Ağaç modelleri her feature için önem skoru üretir**
  - **Gini Importance, Information Gain vb.**

🧭 Feature Selection Yöntemleri 3 Ana Gruba Ayrılır

⚓ **Avantajlar:**

- **Hızlıdır**
- **Model eğitimi ile aynı anda çalışır**
- **Genellikle optimum performans sağlar**

⚠️ **Dezavantajlar:**

- **Sadece seçilen modelle uyumludur**
- **Bazı modellerde özellik önemleri yanlı olabilir (ör. RF yüksek kardinaliteli özellikleri abartabilir)**

Titanic Örneği:

Random Forest feature importance sıralaması genelde şöyle olur:

1. **Sex**
2. **Fare**
3. **Age**
4. **Pclass**
5. **FamilySize**
6. **Düşükler: Cabin, Ticket → elenebilir.**

Thanks For Listening