Woman in Data Science

Lecture 3

02.12.2025

Dersin Amaçları:

- Veri setini okuma ve temel inceleme yapabilecek
- Eksik değerleri tespit edip doldurabilecek
- Outlier (aykırı değer) analizini uygulayabilecek
- Feature selection mantığını anlayacak (teori + pratik)
- Model kurmaya ilk adımı atacak (gelecek ders için hazırlık)

Bölüm — Veri Tanıma ve İlk Gözlem

Titanic - Machine Learning from Disaster | Kaggle

Bölüm — Veri Tanıma ve İlk Gözlem

İlk sorular :

- Hangi değişkenler sayısal, hangileri kategorik?
- Eksik değer en çok nerede?
- Hedef değişkenimiz nedir?

Bölüm — Eksik Değer Analizi

📍 **Amaç → Data-cleaning pratiği**

Bölüm — Eksik Değer Analizi

Uygulama:

- **Age → medyan ile doldurma**
- **Embarked → mod ile doldurma**

Bölüm — Eksik Değer Analizi

🔍 1) %77 missing, feature kalitesi düşüktür

Bir feature'ın %70+ oranında eksik olması, o özelliğin modele katkı sağlayacak kadar güvenilir bilgi taşıyamadığını gösterir.

Bu kadar eksik veriyi doldurmak (impute etmek):

- Büyük ölçüde tahmine dayalı doldurma gerektirir.
- Noise (gürültü) ekleyerek modeli bozabilir.
- Overfitting'e sebep olabilir.

Outlier Analizi (Aykırı Değer)

Her Pclass'ın outlier dağılımına baktığınızda ne görüyorsunuz?

– 1. sınıf / 2. sınıf / 3. sınıf ücret dağılımları nasıl farklı?

– Daha fazla outlier hangi sınıfta?

Outlier Analizi (Aykırı Değer)

Bu farklılık bize gemideki sosyal sınıf yapısı hakkında ne söylüyor?

– Pclass yalnızca bir numara mı?

– Yoksa ekonomik statüyü temsil eden gerçek bir değişken mi?

Outlier Analizi (Aykırı Değer)

Outlier’lar hatalı veri mi yoksa gerçekten var olan ekstrem ücretler mi?

– Örneğin 512 USD gibi aşırı ücretler var → gerçek mi?

– Titanic'te VIP kamaralar gerçekten çok pahalıydı

Outlier Analizi (Aykırı Değer)

Bu outlier'lar hayatta kalma oranını etkileyebilir mi?

– Çok yüksek ücret ödeyenler → daha iyi kamaralar → cankurtaran botlarına daha yakın?

– Yani Pclass + Fare, "Survived" ile ilişkili olabilir mi?

Outlier Analizi (Aykırı Değer)

Outlierları silmek doğru olur mu?

– 1. sınıfta pahalı kabinlere sahip yolcuları silersek...

– Model çok kritik bir bilgiyi kaybeder mi?

Outlier Analizi (Aykırı Değer)

Her Pclass için aynı outlier sınırını kullanmak mantıklı mı?

– 3. sınıfta 20 USD aşırı bir ücret sayılabilir,

– Ama 1. sınıfta 20 USD çok düşük bile olabilir.

Outlier Analizi (Aykırı Değer)

Fare değişkenini modelde nasıl kullanmalıyız?

Bu çok önemli bir yönlendirme:

- Olduğu gibi mi bırakılmalı?
- Log dönüşüm yapmak daha mantıklı mı?
- Pclass + Fare birleştirilerek yeni bir özellik (feature engineering) yapılabilir mi?

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.

📌 Özet

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

Gelecek dersin uygulaması için teorik kısım

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