7 Mart 2023

Açıklanabilir Yapay Zeka

1. Hafta: Giriş

Mustafa Cavus, Ph.D.
🏢 Eskişehir Teknik Üniversitesi - İstatistik Bölümü
📧 mustafacavus@eskisehir.edu.tr
🌐 linktr.ee/mustafacavus

Giriş

Giriş

- Tahmin modelleri insanlık tarihi kadar eskidir. Eski Mısırlılar, Nil nehrinin taşacağını tahmin etmek için Sirius yıldızının hareketlerini gözlemlerdi.

- Model kurmaya dair ilk bilimsel çalışmaların Legendre (1805) ve Gauss (1809) tarafından yayınlanan “En Küçük Kareler Yöntemi”dir.

- Modellerin zamanla ekonomi, tıp, biyoloji ve tarım alanlarında kullanımı arttı.

Giriş

Son yüzyılda tahmin amaçlı kullanılan bir çok istatistiksel model geliştirildi:

- Doğrusal modeller
- Genelleştirilmiş doğrusal modeller
- Sınıflama ve regresyon ağaçları
- Kural tabanlı modeller vb.

Günümüzde bilgisayarların hesaplama gücü ve büyük veri setlerinin kullanılabilirliğindeki artış ile tahmin modellerinin gelişimi hızlanmıştır.

Giriş

- Tahmin modellerine olan talebin artışıyla birlikte modelin esnekliği, değişken seçimi, değişken mühendisliği ve yüksek doğruluk oranına sahip tahmin performansı gibi konular da ilgi görmeye başladılar.

- Böylece *bagging, boosting, stacking* gibi bir çok farklı modelden süper modeller elde etmeye yarayan yöntemlerin kullanımı hız kazandı.

- Bu durum, daha karmaşık modellerin ortaya çıkmasına neden oldu.

Giriş

- Yüksek performanslı karmaşık modellerin en önemli dezavantajı ise *black-box* olarak çalışıyor olmalarıdır.
- Bu tür modellerde değişkenlerin model tahminini nasıl etkilediğini belirlemek zor, hatta imkansız olabilir.
- Bazen bu modelleri beklediğimiz kadar yüksek performans da vermeyebilirler.

Karmaşık modellerin günlük hayatta ortaya çıkardığı sorunlar

Giriş

Bu tür modellerin gerçek hayatta neden olduğu problemler:

- IBM'in *Watson for Oncology* ürünü onkologlar tarafından güvenilir olmayan ve yanlış öneriler sunduğu gibi eleştiriler aldı.
- Amazon'un özgeçmiş tarama sisteminin kadın adaylara karşı önyargılı olduğu tespit edildi.
- Suçluların tekrar suç işleme ihtimalini tahmin etmek için geliştirilen COMPAS algoritmasının Afro-Amerikalılara karşı önyargılı olduğu görüldü.
- Apple ödeme kartlarının arkasında yer alan algoritmaların farklı cinsiyetlere karşı farklı çalıştığı ortaya çıktı.
- Veri kaymalarının model performansında bozulmaya yol açtığı duruma örnek olarak, lansmanından iki yıl sonra dah kötü tahminler veren Google Flu modeli örnek verilebilir.

Yasal düzenlemeler

Yasal düzenlemeler

Bu sorunlar ile karşılaşmamak için bazı önemler alındı:

- ***General Data Protection Regulations* (GDPR 2018)**
- ***Right to explanation***, bir algoritmanın çıktısının açıklanması hakkıdır.
- ***The AI Act (EU 2022)***, AI uygulamarını üç risk kategorisine ayırdı ve denetim altına aldı.

Giriş

Bu sorunlar ve alınan önlemler,

- Alan bilgisi
- Model seçimi
- Model doğrulama

gibi konuların önemini daha da arttırdı.

Giriş

Özetle günümüzde tahmin modellerinde karşılaşılan sorunlar, (1) veri yetersizliği, (2) hesaplama gücü eksikliği, (3) yetersiz algoritmalar, ve (4) esnek modeller değil,

- **Modelin keşfi ve açıklanması** (model tahminlerinin nasıl gerçekleştiğinin anlaşılması)
- **Modelin incelenmesi** (model performansının yetersiz olduğu noktaların belirlenmesi)

için kullanılabilecek araçların eksikliğidir.

Motivasyon

Motivasyon

Bu gibi nedenlerden dolayı, modelin işleyişini (iç yapısını) anlamaya yönelik olarak ortaya çıkan ve birbirleri yerine sıklıkla kullanılan terimler ile adlandırılan **Açıklanabilir yapay zeka**, **açıklanabilir makine öğrenmesi**, **yorumlanabilir makine öğrenmesi** ya da **açıklayıcı model analizi** aktif bir araştırma alanıdır. Her geçen gün çok sayıda çözüm geliştirilmeye devam etmektedir.

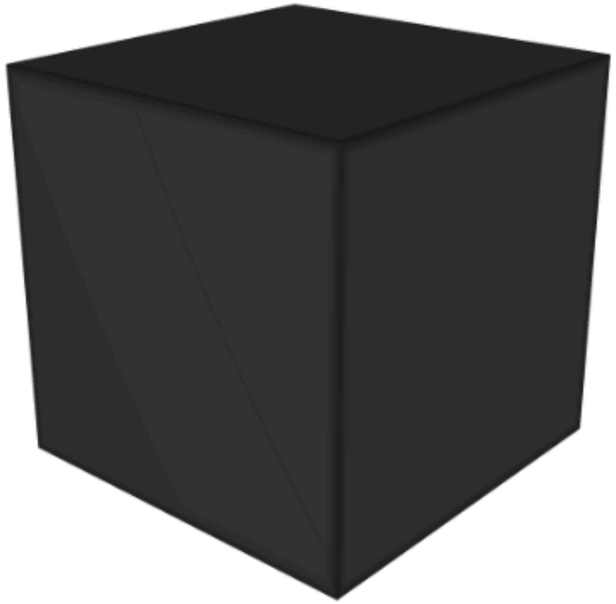
Bu derste hedeflenen model keşfinde kullanılan temel kavram ve yöntemleri tanımak ve bu süreçte karşılaşılan sorunların çözümleri için kullanılan yöntemleri kullanabilmektir.

Model yapıları

Black-box vs. Glass-box

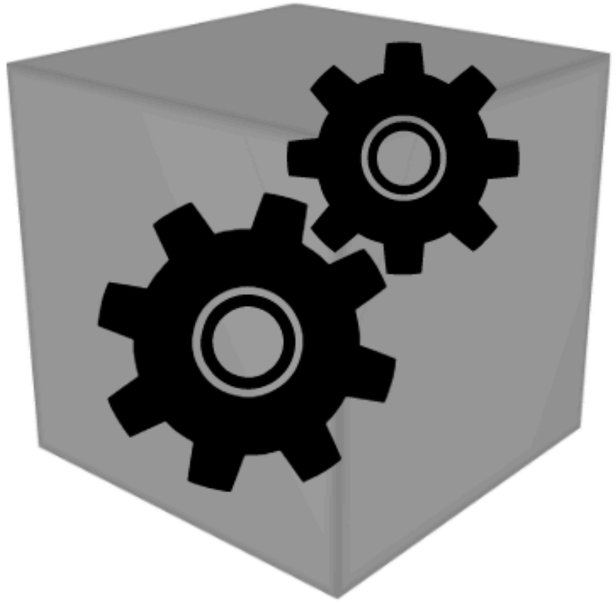
Genellikle *black-box* modeli terimi, insanlar tarafından anlaşılması zor olan karmaşık bir yapıya sahip modeller için kullanılmaktadır.

Bu genellikle çok sayıda model katsayılarına veya karmaşık matematiksel dönüşümlere atıfta bulunur.



Black-box

Opaque



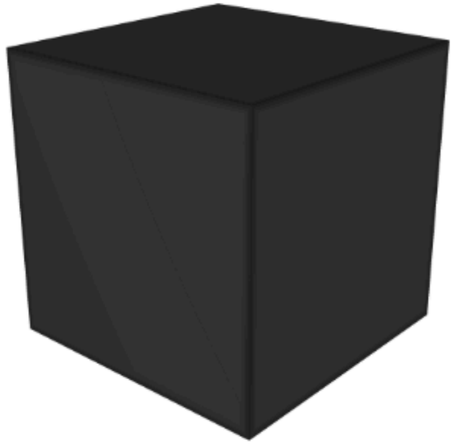
Glass-box

White-box

Transparent

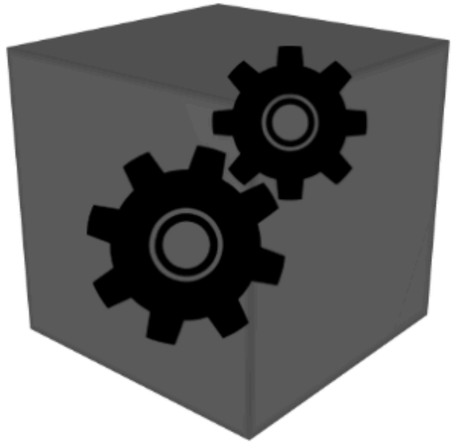
Black-box vs. Grey-box vs. Glass-box

Boosted trees



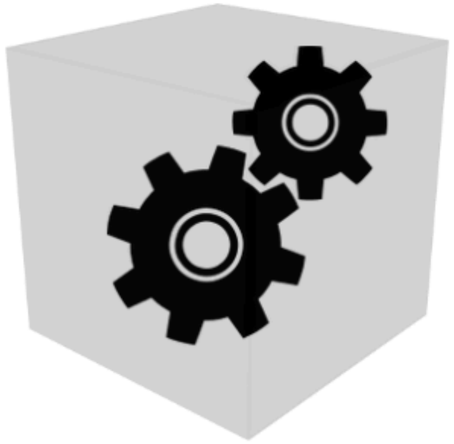
Black-box

Random forest



Gray-box

Decision tree

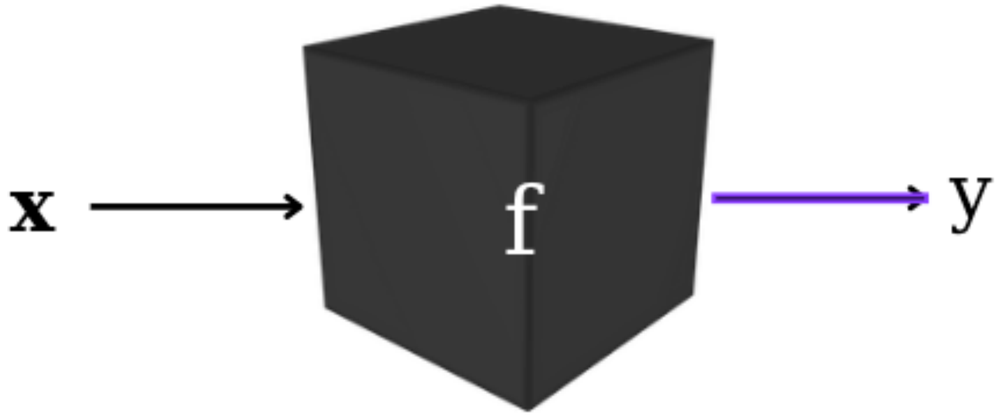


Glass-box

Açıklayıcılar

Açıklayıcılar

Kara-kutu yapısındaki modellerin açıklanması için, Açıklanabilir Yapay Zeka (*Explainable Artificial Intelligence*) araçları kullanılır. Bu ders boyunca bu araçlardan kısaca **açıklayıcılar** olarak bahsedilecektir.



Açıklayıcı türleri

Açıklayıcılar

Açıklayıcılar, model bağımlı ve bağımsız olmaları, lokal (gözlem düzeyinde) ya da global (veri seti düzeyinde) olarak kullanılmasına göre sınıflandırılırlar.

Modelden bağımsız ve modele bağlı açıklayıcılar

Modelden bağımsız yaklaşımlar tüm modeller ile kullanılabilirken modele bağlı yaklaşımlar yalnızca bağlı olarak geliştirildikleri modeller ile kullanılabilirler. Bu nedenle modelden bağımsız yaklaşımların pratikte kullanım alanı daha fazladır. Buna karşılık, bazı durumlarda modelden bağımsız yaklaşımların iyi performans göstermedikleri bilinmektedir.

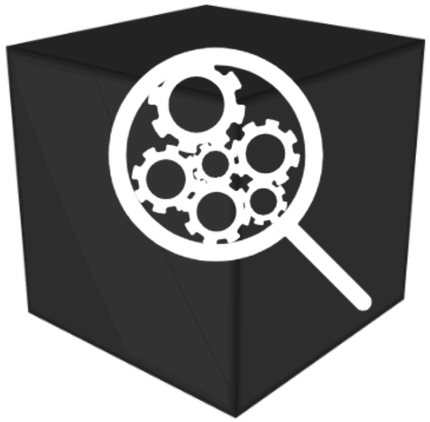
Modelden bağımsız

Modele bağlı

Lokal ve Global düzeyde açıklayıcılar

Lokal düzeydeki açıklayıcılar, modeli bir gözlem düzeyinde açıklamak için kullanılırken, global düzeydeki açıklayıcılar modeli tüm gözlemler düzeyinde bir bütün olarak açıklamak için kullanılırlar.

Lokal
Gözlem
Tahmin

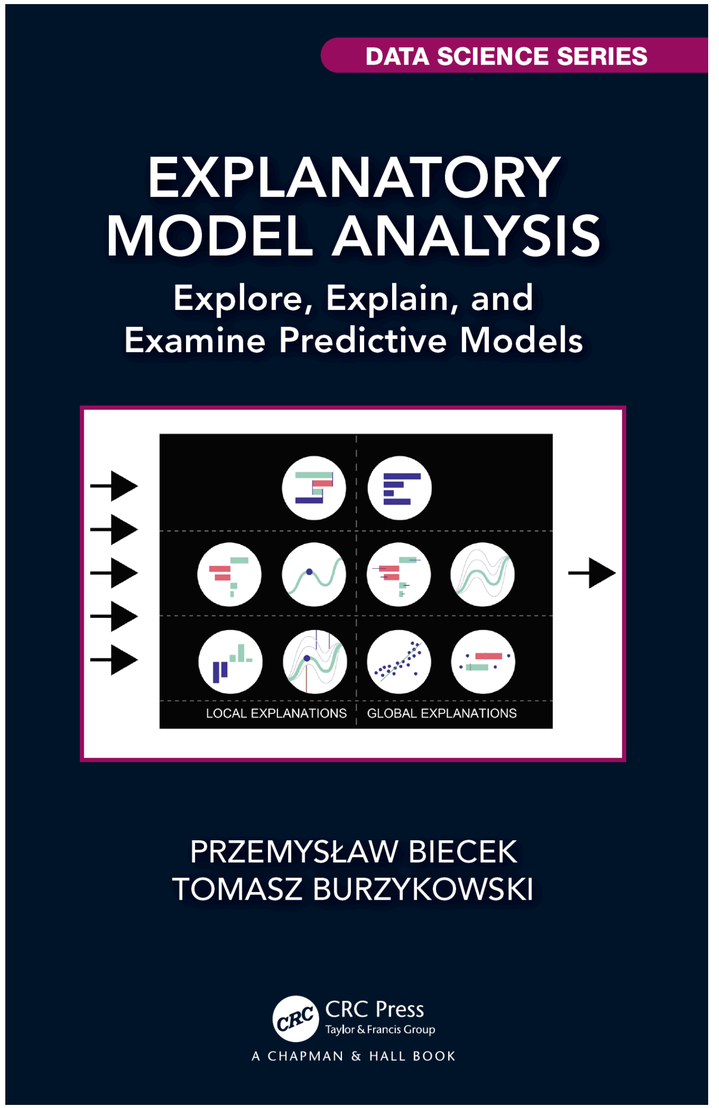


Global
Veri seti
Model

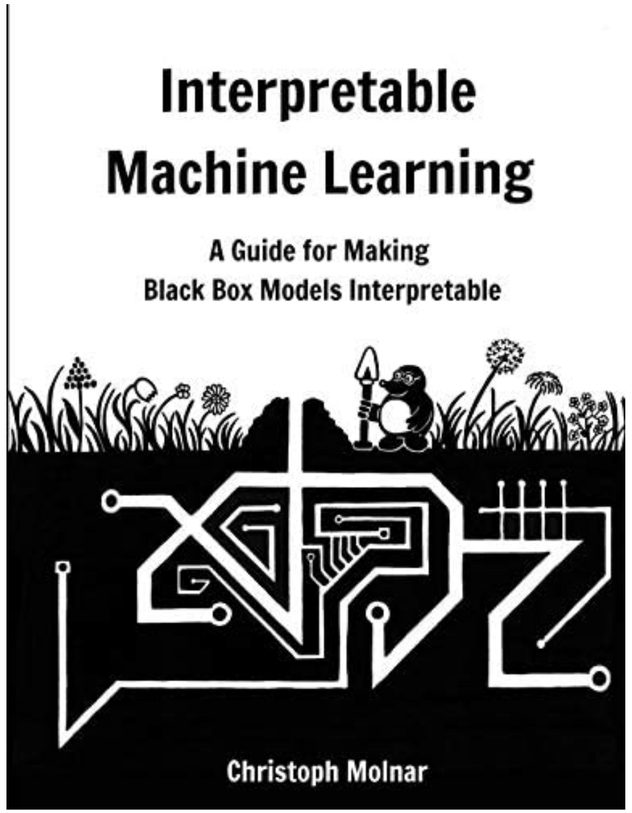
Son yıllarda, bazı uygulama alanlarında ortaya çıkan ihtiyaçlar nedeniyle birçok sayıda lokal açıklamaların birleştirilmesi yaklaşımına dayalı olan, **Glokal** düzeyde açıklayıcılar kullanılmaya başlamıştır.

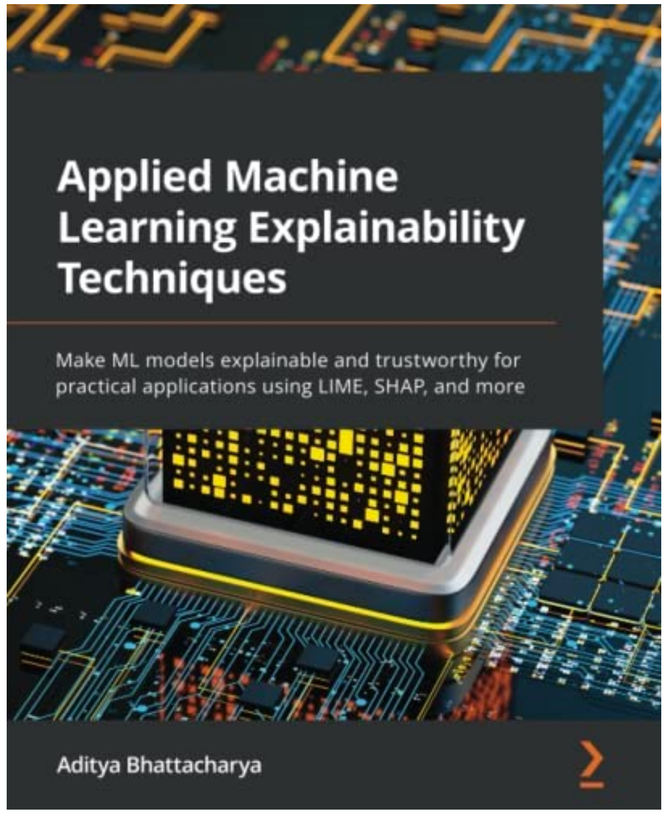
Kaynaklar

Ders materyallerinin hazırlanmasında **Explanatory Model Analysis (Biecek and Burzykowski, 2021)** kitabından yararlanılmıştır. Kitabın ücretsiz online versiyonuna bağlantı üzerinden erişilebilir: **https://ema.drwhy.ai/**



Öneriler





İleri okumalar için,
Interpretable Machine Learning (Molnar, 2023) ve **Applied Machine Learning Explainability Techniques (Bhattacharya, 2022)** kitapları önerilir.

https://christophm.github.io/interpretable-ml-book/

Ders notlarına dersin **GitHub** sayfası üzerinden ulaşabilirsiniz.

Ders ile ilgili sorularınız için mustafacavus@eskisehir.edu.tr adresi üzerinden benimle iletişime geçebilirsiniz.

Mustafa Cavus, Ph.D.
🏢 Eskişehir Teknik Üniversitesi - İstatistik Bölümü
📧 mustafacavus@eskisehir.edu.tr
🌐 linktr.ee/mustafacavus