28 Mart 2023

Açıklanabilir Yapay Zeka

4. Hafta: Lokal düzeyde açıklayıcılar | LIME yöntemi

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

Giriş

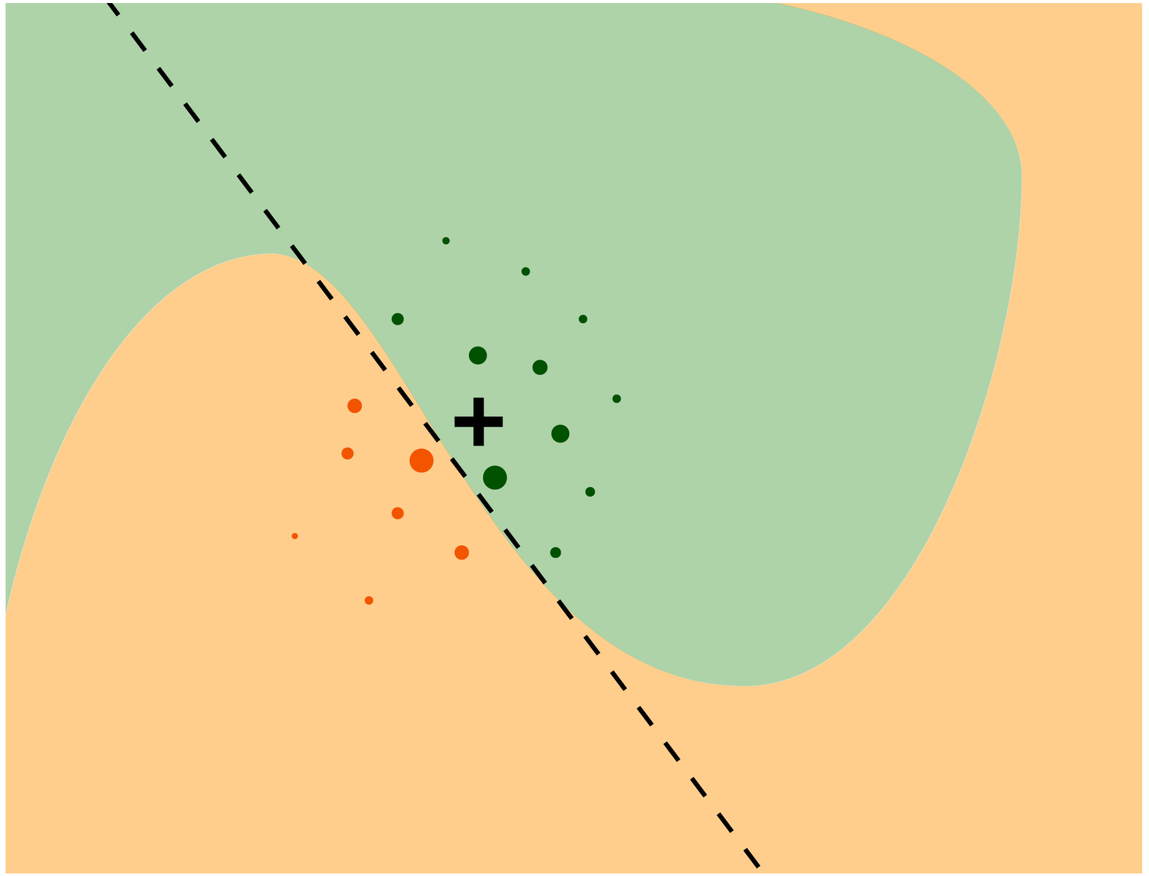
- Break-Down ve SHAP gibi yöntemler çok sayıda değişken sayısına sahip modeller için uygun değildir.

- Ancak gerçek hayatta kullanılan modellerde yüzlerce hatta bincelerce açıklayıcı değişken içeren modeller kullanılmaktadır.

- Bu gibi durumlarda az sayıda değişken içeren açıklayıcılar iyi bir alternatiftir. Bunlardan en bilineni Yerel Yorumlanabilir Modelden Bağımsız Açıklamalar (**LIME**: **L**ocal **I**nterpretable **M**odel agnostic **E**xplanations)

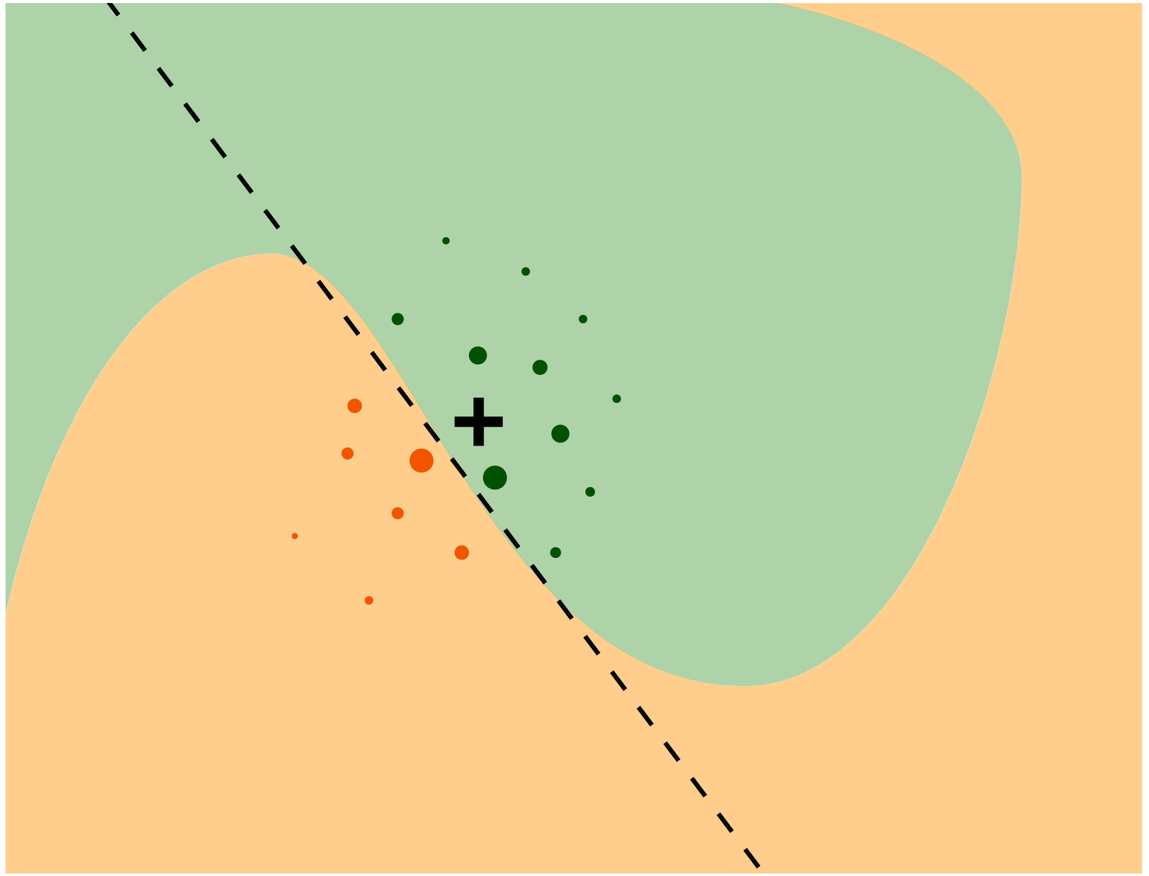
LIME Yöntemi

- Ribeiro vd. (2016) tarafından önerilmiştir.
- LIME yöntemi, yorumlanması daha kolay olan daha basit bir *glass-box* modeli ile bir *black-box* modeline yerel olarak yakınsama fikrine dayanır.



LIME Yöntemi

- Renkli alanlar, karmaşık bir ikili sınıflandırma modeli için karar bölgelerine karşılık gelir.
- Siyah çarpı ilgilenilen gözlem değerini temsil eder.
- Noktalar, ilgilenilen örnek etrafındaki yapay verilere karşılık gelir.
- Kesikli çizgi, yapay verilere uyan basit bir doğrusal modeli temsil eder. Basit model, ilgilenilen gözlem değeri etrafındaki *black-box* modelinin yerel davranışını “açıklar".

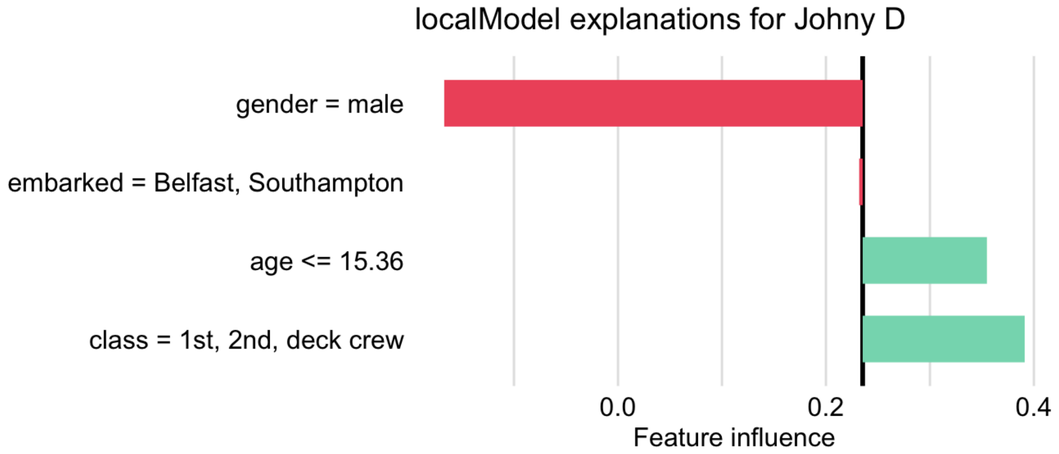


Örnek uygulama

- Titanic veri setinde yer alan bazı değişkenleri, ikili kategorik değişkene dönüştürerek daha basit bir veri yapısı elde edelim.

- Daha sonra ilgili gözlem değeri etrafında rasgele orman modeli ile yapay gözlemler üretelim.

- Ardından, gözlemler üzerinde K-LASSO modelini kullanarak yorumlanabilir yerel bir model elde edelim.



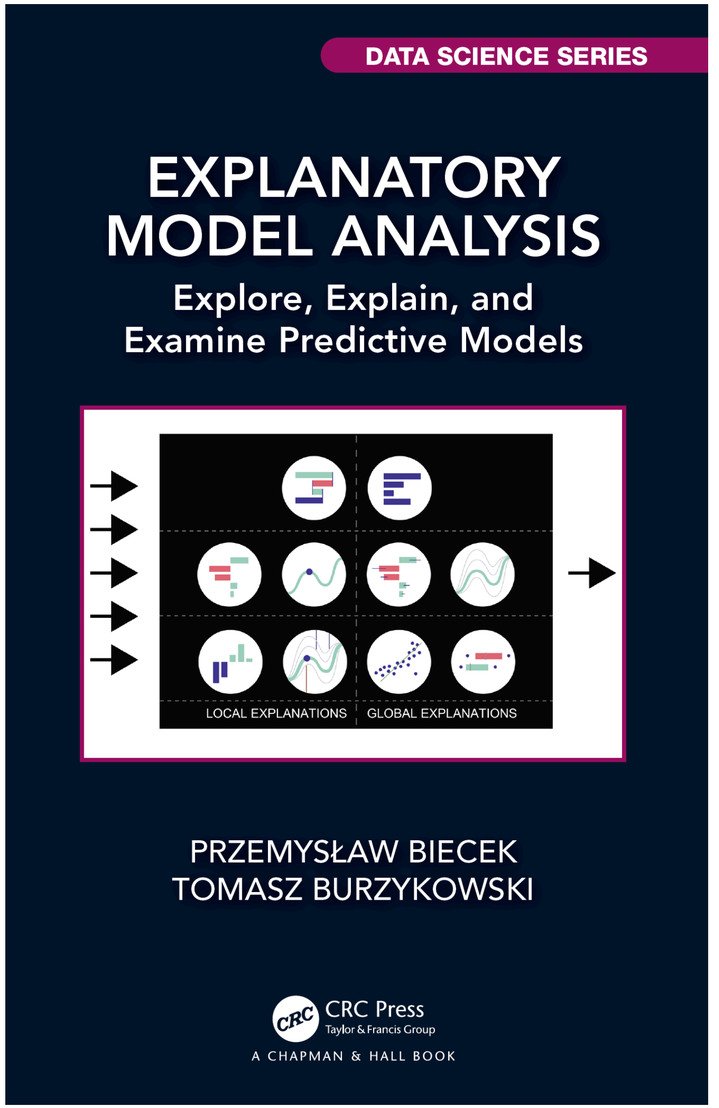
Artı ve eksileri

- Modelden bağımsızdır.
- Yerel doğruluk (*local fidelity*) sağlar. Yani black-box modele yerel düzeyde uyumludur.
- Yüksek boyutlu (çok sayıda değişken içeren) modellere uygulanabilir.

- Açıklamanın doğruluğu yorumlanabilir modelin seçimine bağlıdır. Açıklayıcı bunu kontrol etmez.
- Yüksek boyutlu verilerde gözlemlerin yerel komşularını üretmek kolay olmayabilir. Komşu noktalardaki küçük değişiklikler açıklamaları etkileyebilir.

Kaynaklar

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



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