14 Mart 2023

Açıklanabilir Yapay Zeka

2. Hafta: Lokal düzeyde açıklayıcılar | Break-Down yöntemi

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

Giriş

Lokal açıklamalar

Lokal açıklamalar, bir modelin belirli bir gözlem değeri için nasıl bir tahmin sağladığını anlamamıza yardımcı olur. Genellikle üç temel soruya yanıt aramak için başvurulur:

1. Açıklayıcı değişkenlerin modelin tahminleri üzerindeki etkilerini değerlendirmek

2. Bazı açıklayıcı değişkenlerin değerleri değişirse modelin tahminlerinin nasıl değişeceğini anlamak

3. Modelin yanlış tahminler sağladığını keşfetmek ve nedenini bulmak

Lokal açıklayıcılar

Lokal düzeydeki açıklayıcılar, bir modeli ilgilenilen bir gözlem düzeyinde açıklamak için kullanılan araçlardır. Bu tür araçlar üç gruba ayrılırlar:

1. Değişken katkılarının (*variable attributions*) analizine dayalı açıklayıcılar
2. Bir gözlem değeri etrafında model davranışının analizine dayalı yöntemler
3. "*What-if*" analizine dayalı açıklayıcılar

1. Değişken katkılarının analizine dayalı açıklayıcılar

Bir yaklaşım, modelin belirli bir örnek için tahmininin ortalama tahminden ne kadar farklı olduğunu ve farkın açıklayıcı değişkenler arasında nasıl dağıldığını analiz etmektir.

1. Break-Down yöntemi
2. Etkileşimler için Break-Down yöntemi
3. Shapley toplamsal açıklamalar (SHAP) yöntemi

2. Bir gözlem değeri etrafında model davranışının analizine dayalı yöntemler

Modelin bir fonksiyon olarak yorumlanmasına dayalı olarak ilgilenilen bir gözlem değeri etrafındaki yerel davranışını araştıran yöntemlerdir. Özellikle, model yanıt (tahmin) yüzeyinin gözlem değeri etrafındaki davranışı analiz edilir. **Bir *black-box* modelin, ilgili gözlem değeri etrafındaki davranışı, daha basit bir *glass-box* modeliyle taklit edilebilir.**

1. LIME yöntemi

3. “What-if” analizine dayalı açıklayıcılar

Tek bir açıklayıcı değişkenin değeri değişirse modelin tahmininin nasıl değiştiğini araştıran açıklayıcılardır. Bu yaklaşım, “What-if” analizlerinde kullanışlıdır. Özellikle, tek bir açıklayıcı değişkendeki değişikliğin neden olduğu model tabanlı tahminlerdeki değişikliği gösteren grafikler oluşturulabilir.

1. Ceteris-paribus profilleri
2. Ceteris-Paribus salınımları
3. Yerel tespit eğrileri

Break-down yöntemi

Break-Down yöntemi

Modelin bir gözlem için tahminini anlamaya çalışırken muhtemelen en sık sorulan soru şudur:
“Bu sonuca en çok hangi değişkenler katkıda bulundu?”

- Toplamsal etkiler içeren modellerin gözlem düzeyinde davranışını açıklamak için geliştirilmiştir.
- Bir gözlem için tahmin edilen yanıt değişkeni değerine, her bir değişkenin sunduğu katkıyı analiz etmek için kullanılır.

Break-Down (BD) yöntemi

Burada $f(\underline{x})$ tahmininin, $\underline{x}$ değerleri verildiğinde bağımlı değişken Y'nin beklenen değerinin bir tahmini olduğunu varsayalım:

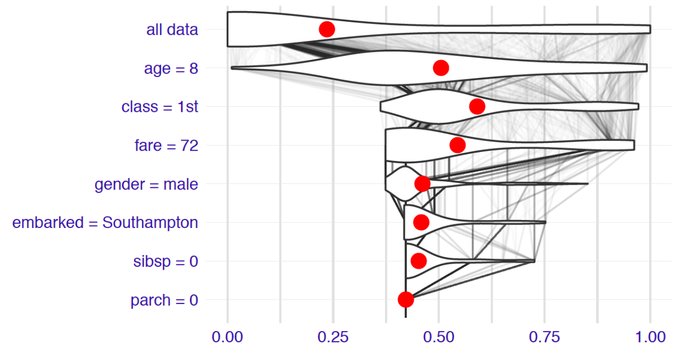
BD yöntemi, diğer değişkenlerin değerlerini sabitlerken Y'nin beklenen değerindeki değişimi hesaplayarak açıklayıcı bir değişkenin modelin tahminine katkısını ölçme fikrine dayalıdır.

Break-Down yöntemi

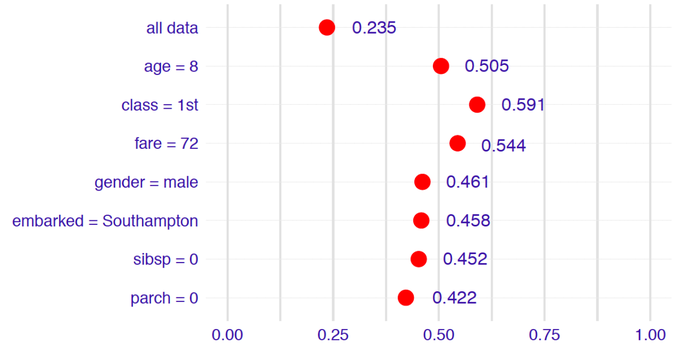
İlgilenilen bir gözlem değeri için BD grafiklerinin oluşumunu üç aşamada inceleyelim:

A) Burada yer alan ilk keman grafiği, modelin tüm tahmin değerleri üzerinden oluşturulur. İkinci sırada yer alan ***age*** değişkeni için çizilen keman grafiği, tüm tahmin değerleri arasında ***age*** = 8 olanlar ile oluşturulmuştur. Üçüncü sırada yer alan, ***class*** değişkenine karşılık gelen grafik ise hem ***age*** = 8 hem ***class*** = 1st olan tahmin değerleri üzerinden elde edilir. Bu işlemler kırılımlı bir şekilde tüm değişkenler boyunca devam eder.

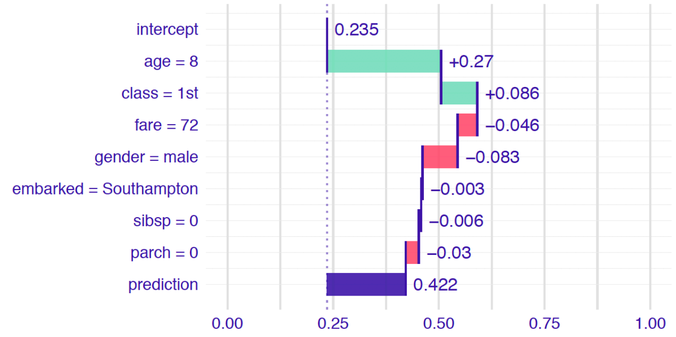
A)



B)

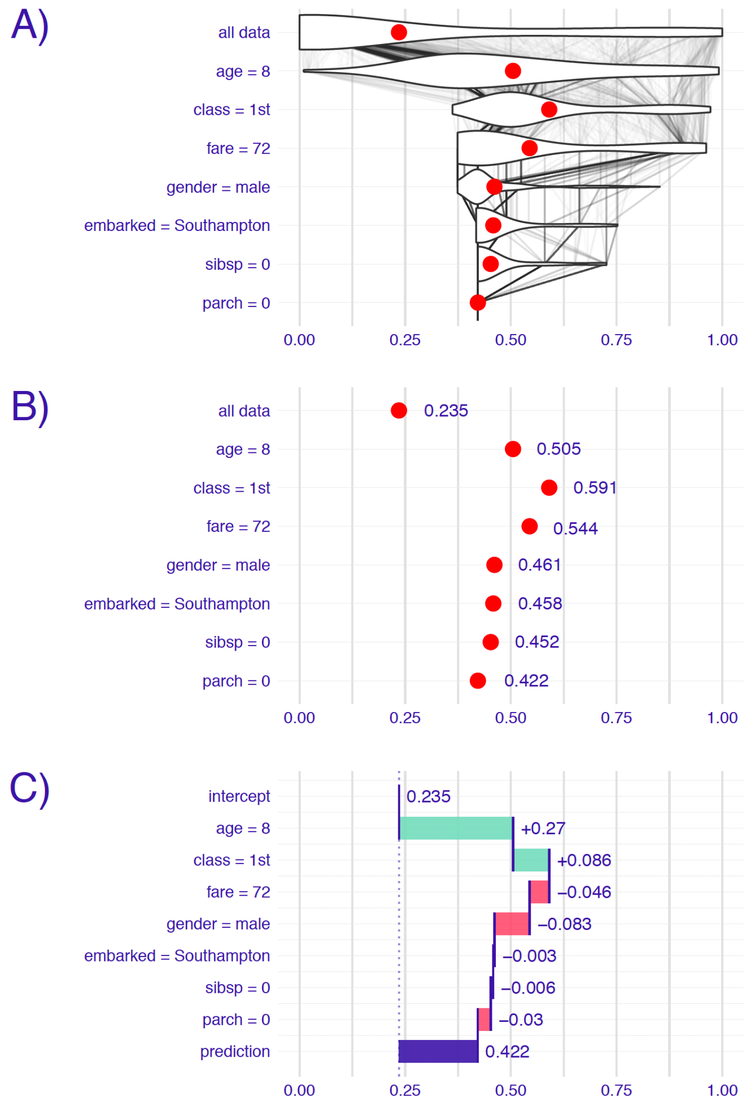


C)



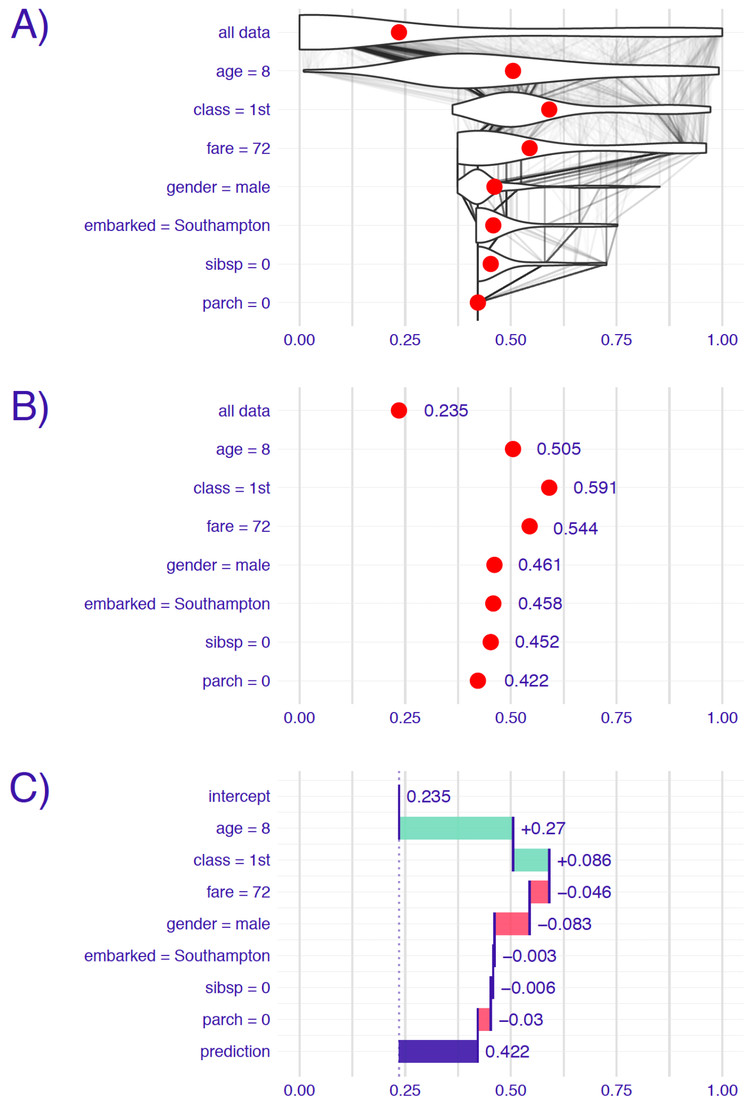
Break-Down yöntemi

B) Bir önceki aşamada keman grafiklerini oluşturmak için kullanılan tahmin değerlerinin ortalamaları kırmızı noktalar ile işaretlenmiştir. İkinci aşamada kırmızı noktalar tek başına kalacak şekilde keman grafikleri ortadan kaldırılır.



Break-Down yöntemi

C) Son aşamada, ilk sıradaki kesim noktası (*intercept*) başlangıç noktası olarak alınmak üzere, izleyen değerler eğer bir öncekinden daha büyükse pozitif katkı, daha küçükse negatif katkı sunduklarını belirtmek için yeşil ve kırmızı renk ile belirtilirler. Son olarak başlangıç noktasından, tahmin değerini (*prediction)* göstermek üzere mor bir çubuk ile grafik tamamlanır.



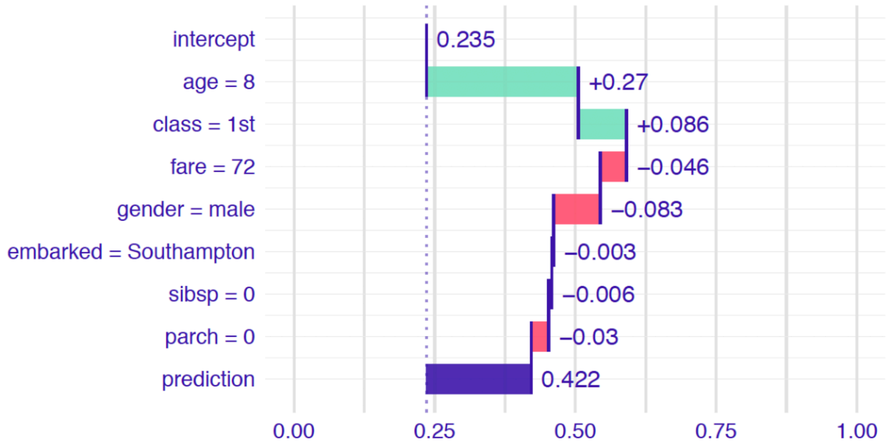
Break-Down yöntemi

Çizilen BD grafiği üzerinden, y-ekseninde her bir değişkenin aldığı değerleri verilen gözlem değeri üzerinden hesaplanan tahmine değişkenlerin katkısı incelenebilir.

En yüksek pozitif katkıyı **age = 8**, negatif katkıyı ise **gender = male** değişkenleri yapmıştır.

Burada altı çizilmesi gereken nokta, değişkenlerin katkısının aldıkları değerlere de bağlı olmasıdır.

C)

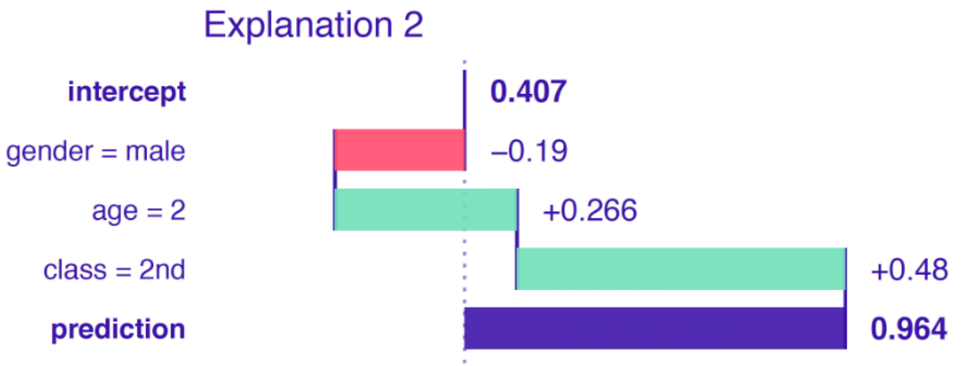


BD yöntemi toplamsal etkilere sahip (değişkenler arasında etkileşim olmayan) modeller üzerinde iyi çalışırlar. Etkileşim etkisi olduğunda, hesaplanan her bir değişkenin katkı düzeyi ve yönü değişkenlerin sırasına bağlı olarak değişkenlik gösterebilir.

Break-Down yöntemi

Etkileşim durumunda BD yöntemi ile değişken katkılarını incelemenin sonuçlarını bir örnek üzerinde inceleyelim.

Yandaki iki grafik aynı gözlem değeri için ***gender***, ***age*** ve ***class*** değişkenlerinin yerleri değiştirilerek elde edilen iki lokal açıklamayı göstermektedir.

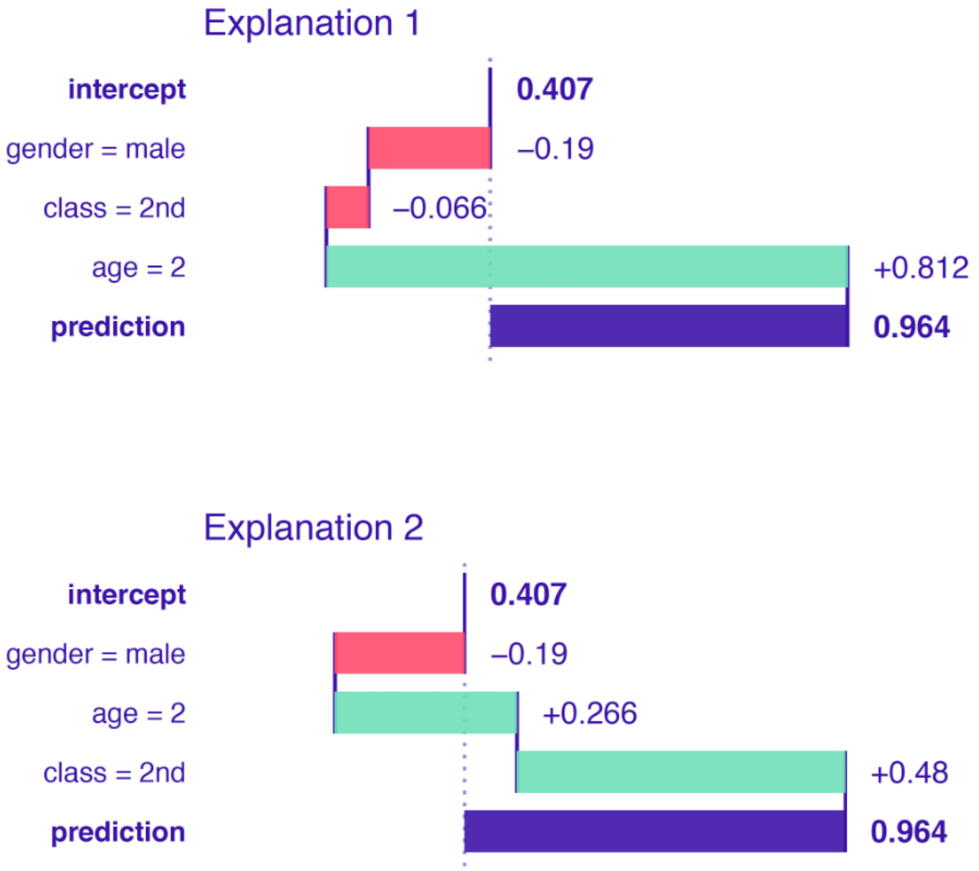


Break-Down yöntemi

Explanation 1 grafiğinde, model tahminine ilk iki değişkenin negatif, üçüncü değişken ise pozitif katkısı olduğu görülmektedir.

Yolcunun erkek olması, ortalama model tahminine kıyasla hayatta kalma şansını azaltır. Hayatta kalma olasılığını daha da azaltan ikinci sınıfta seyahat etmesidir.

Ancak, yolcunun yaşı çok küçük olduğu için hayatta kalma şansını önemli ölçüde artırıyor. Bu sonuç, ikinci sınıftaki yolcuların çoğunun yetişkin olmasının sonucudur; bu nedenle, ikinci sınıftan bir çocuğun hayatta kalma şansı daha yüksektir.



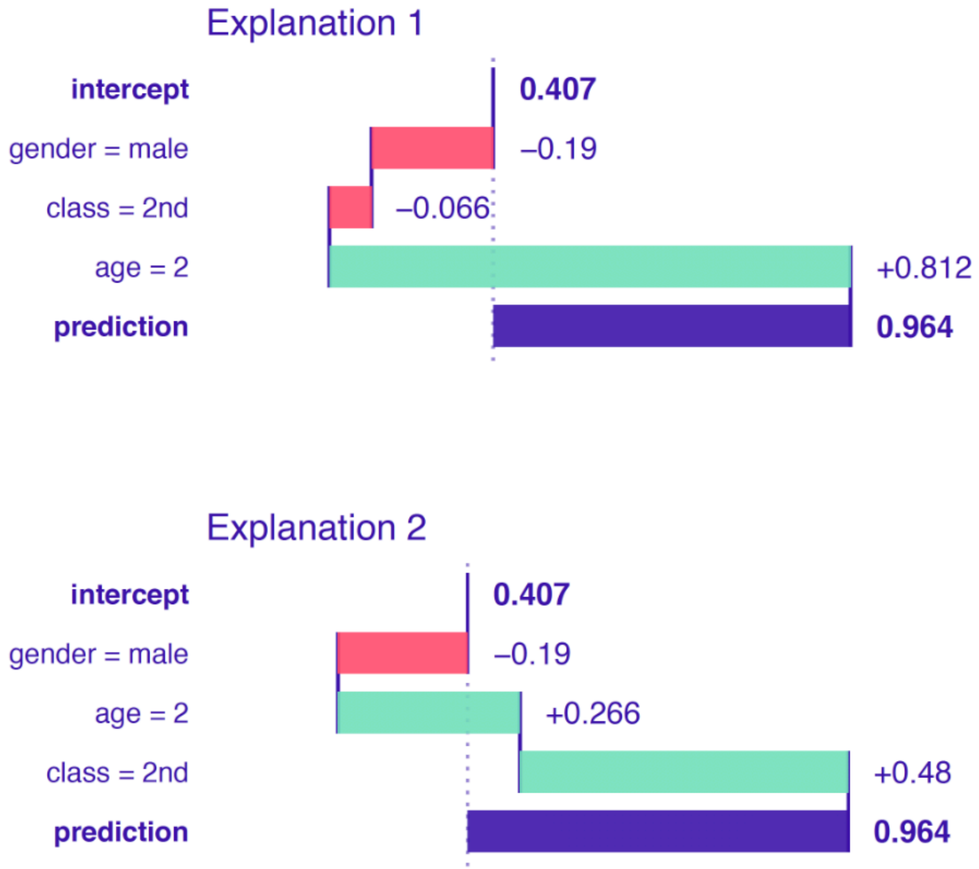
Break-Down yöntemi

Explanation 2 grafiğinde, açıklayıcı değişkenlerin sırası: ***gender***, ***age*** ve ***class*** olarak değiştirilmiştir.

Bir önceki grafikte yer alan açıklamadan farklı olarak sınıfın pozitif katkısı olduğu görüşmektedir.

Yolcunun erkek olması, ortalama model tahminine kıyasla hayatta kalma şansını azaltır. Ancak çok genç olması, yetişkin erkeklere kıyasla hayatta kalma olasılığını artırıyor. Son olarak çocuğun ikinci sınıfta seyahat etmesi hayatta kalma şansını daha da artırıyor.

Bu sonuç, çoğu çocuğun üçüncü sınıfta seyahat etmesinden kaynaklanmaktadır; bu nedenle ikinci sınıfta yolculuk eden bir çocuk olmak hayatta kalma şansını artırmaktadır.



Artı ve eksileri

- Modelden bağımsız
- Görselleştirilebilir
- Anlaşılması kolay
- Kompakt

- Değişkenler arası etkileşime karşı dayanaklı değil
- Çok sayıda açıklayıcı değişken olması durumunda işlevsizleşebilir (*complexity*)

Etkileşim durumunda
Break-down yöntemi

Etkileşim durumunda Break-Down yöntemi

Etkileşim durumunda kullanılan BD yöntemi, değişken çiftleri arasındaki etkileşimleri tespit eder ve grafikleri oluştururken etkileşimleri dikkate alır.

Etkileşim (*deviation from additivity*), bir açıklayıcı değişkenin etkisinin diğer değişkenin ya da değişkenlerin değerlerine bağlı olduğu anlamına gelir.

Etkileşim durumunda Break-Down yöntemi

Örnek olarak, ***titanic*** veri setinde yer alan ***age*** ve ***class*** değişkenlerini ele alalım. Basitleştirmek için sürekli olan ***age*** değişkenini, 0-16 yaş ve 17 yaş ve üzeri olmak üzere, ***class*** değişkenini de 2.sınıf ve diğerleri olmak üzere ikili kategorik değişkenlere dönüştürelim. Aşağıda tablo erkek yolcuların hayatta kalma yüzdelerini göstermektedir.

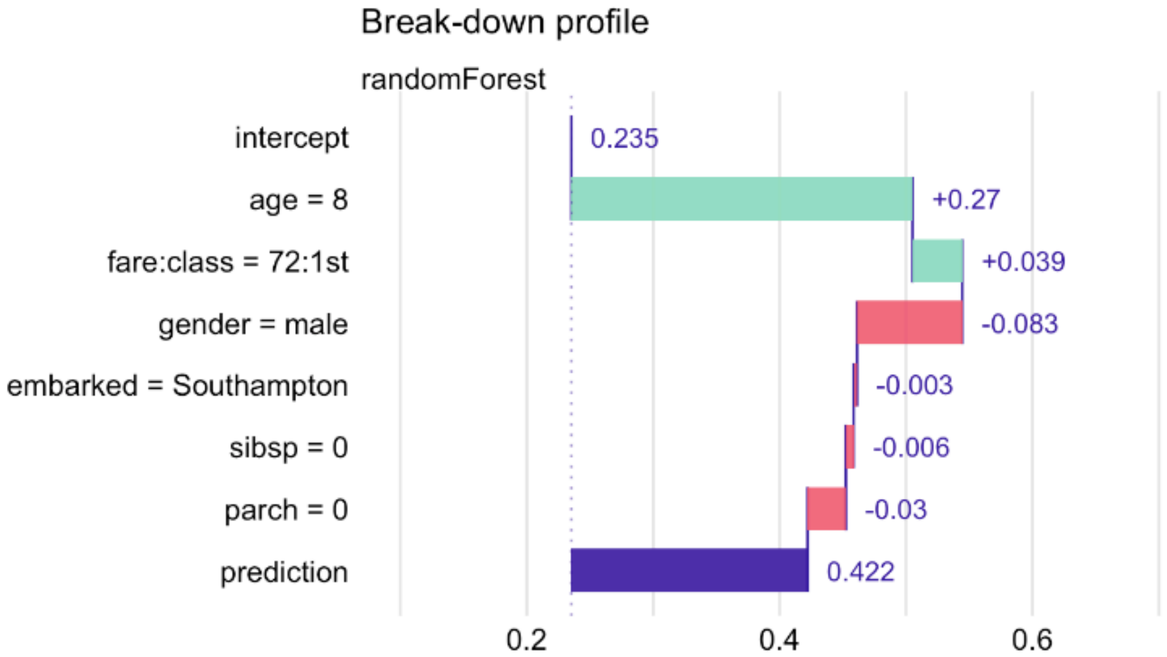
Class	**0-16**	**17**	**Toplam**
2nd	%91.7	%7.8	%13.5
other	%31.9	%20.8	%21.3
Toplam	%40.7	%19.5	%20.5

Etkileşim durumunda Break-Down yöntemi

Bu örnekte etkileşimlerin bir modelin tahminlerine göre açıklayıcı değişkenlerin katkısını değerlendirmeyi zorlaştırdığı görülmektedir.

Etkileşim durumunda Break-Down yöntemi

Olası etkileşim etkileri hesaplandıktan sonra değişkenlerin net etkileri ile arasında fark olması durumunda, etkileşimli değişkenler de dikkate alınır.



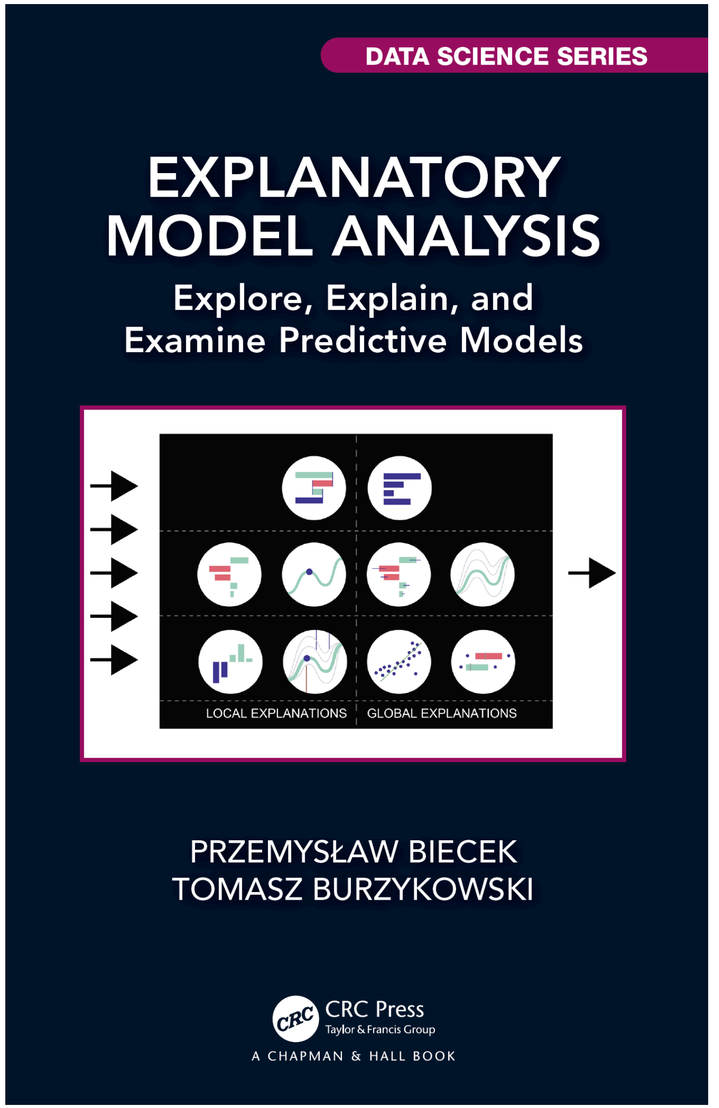
Artı ve eksileri

- Modelden bağımsız
- Görselleştirilebilir
- Anlaşılması kolay

- Etkileşim etkilerinin tespiti herhangi bir istatistiksel yönteme dayalı olmadığı için küçük örneklem hacimlerinde hatalı sonuçlara yol açabilir.
- Etkileşim etkilerini dikkate aldığı için daha fazla hesaplama karmaşıklığına sahiptir.

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