21 Mart 2023

Açıklanabilir Yapay Zeka

3. Hafta: Lokal düzeyde açıklayıcılar | SHAP yöntemi

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

Giriş

Break-down yönteminin en önemli eksikliği, modelde etkileşim etkisi olması durumunda sonuçların değişkenlerin sırası bağlı olmasıdır. Yani Break-down grafiği oluşturulurken kullanılan değişkenlerin sırası değiştiğinde, değişkenlerin katkıları da değişmektedir.

Giriş

SHAP yöntemi, değişkenleri tüm olası sıralamaları üzerinden hesaplanan katkı değerlerinin ortalamasının alınması fikrine dayalıdır.

Giriş

- SHAP yöntemi, Oyun Teorisi alanında Shapley (1953) tarafından geliştirilen “Shapley değerleri” yöntemine dayalı bir yaklaşımdır.
- Yaklaşım ilk olarak Štrumbelj ve Kononenko (2010) ve Štrumbelj ve Kononenko (2014) tarafından ML alanında kullanılmıştır.
- Lundberg (2019) tarafından yazılan SHAP Python kütüphanesi sonrasında oldukça popülerleşmiştir.

Shapley Değerleri

Shapley değerleri aşağıdaki problem ile ilgilenir:

Birden fazla oyuncudan oluşan bir takım belirli bir hedefe ulaşmak için bir oyun oynarlar. Oyuncular farklıdır ve farklı oyuncuların oyuna farklı düzeyde katkıları olabilir. İşbirliği yapmaları faydalıdır, çünkü bireysel eylemlerden daha fazla fayda sağlayabilir. Çözülmesi gereken sorun, üretilen fazlalığın oyuncular arasında nasıl dağıtılacağıdır. Shapley değerleri, bu soruya cevap verir (Shapley 1953).

Shapley Değerleri

Bu sorunu ML perspektifiyle ele alalım:

- Açıklayıcı değişkenler oyuncular, model ise oyunculardan oluşan takım olduğunu düşünelim.
- Takım olmanın (birden fazla oyuncunun bir araya gelmesi) getirisi, modelin tahminidir.
- Çözülmesi gereken problem, modelin tahmininin değişkenlerin sağladığı katkıların belirlenmesidir.

Shapley Değerlerinin Özellikleri

Aşağıdaki özellikleri sayesinde Shapley değerleri kolaylıkla ML alanına entegre edilebilmişlerdir:

1. **Simetriklik**: Eğer modelde yer alan herhangi iki değişken birbirlerinin yerine kullanılabilirlerse (eşdeğer), Shapley değerleri birbirine eşittir.

2. **Etkisizlik**: Eğer modeldeki bir değişkenin model tahminine hiçbir katkısı yok ise Shapley değeri sıfırdır.

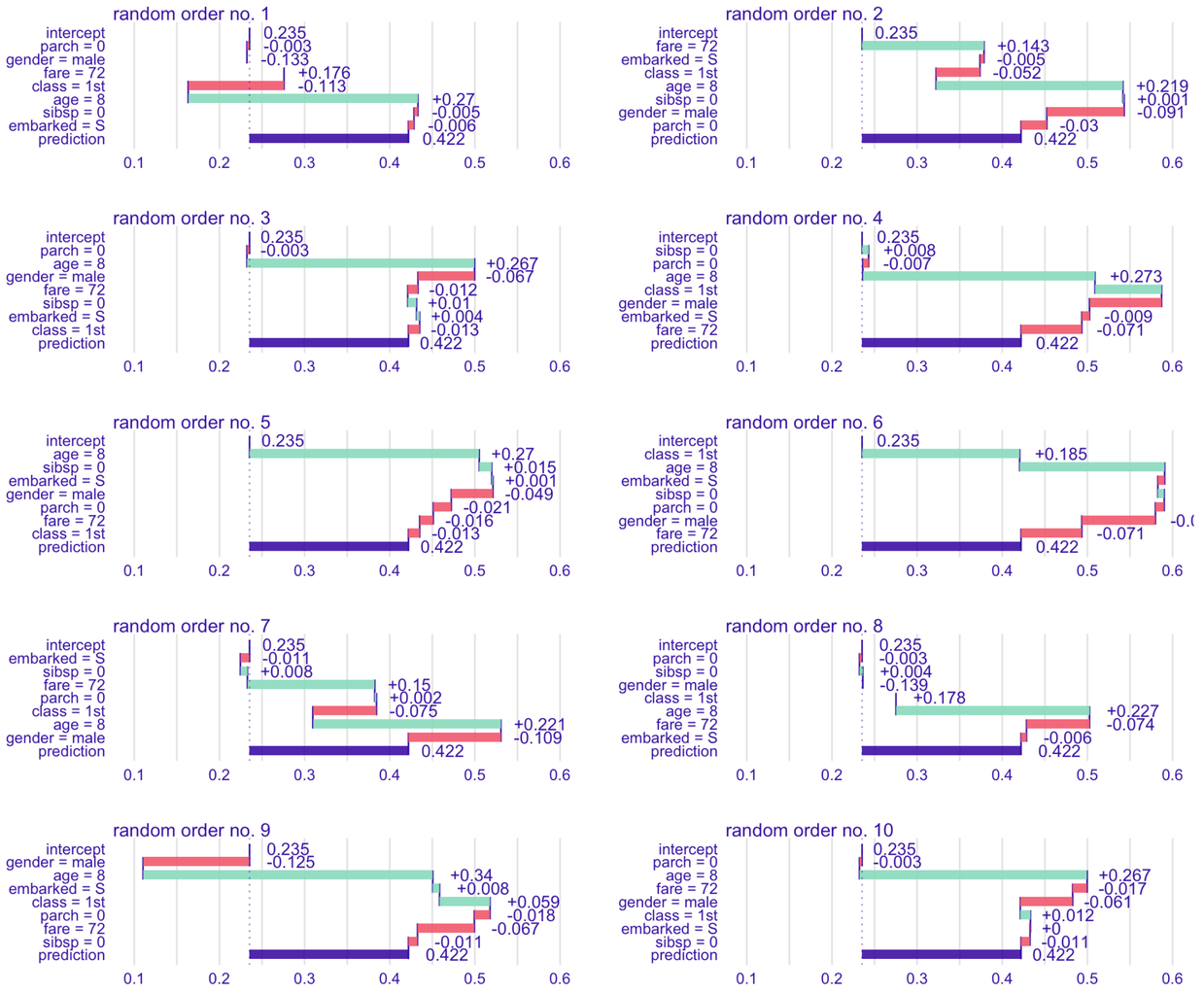
3. **Toplamsallık**: Eğer bir f modeli, g ve h gibi iki modelin toplamı şeklinde ise f modeli için hesaplanan Shapley değeri g ve h modelleri için hesaplanan Shapley değerlerinin toplamına eşittir.

4. **Lokal doğruluk**: Shapley değerlerinin toplamı model tahmin değerine eşittir.

SHAP yöntemi

Yanda *titanic* veri seti üzerinde eğitilen bir *random forest* modelinin farklı değişken sıralamalarına karşılık oluşturulan Break-down grafikleri verilmiştir.

Grafikte farklı değişken sıralamaları için bazı değişkenlerin model tahminine katkılarının farklı oldukları görülmektedir. Öne çıkan farklılıklar ***fare*** ve ***age*** değişkenlerinde gözlenmektedir.

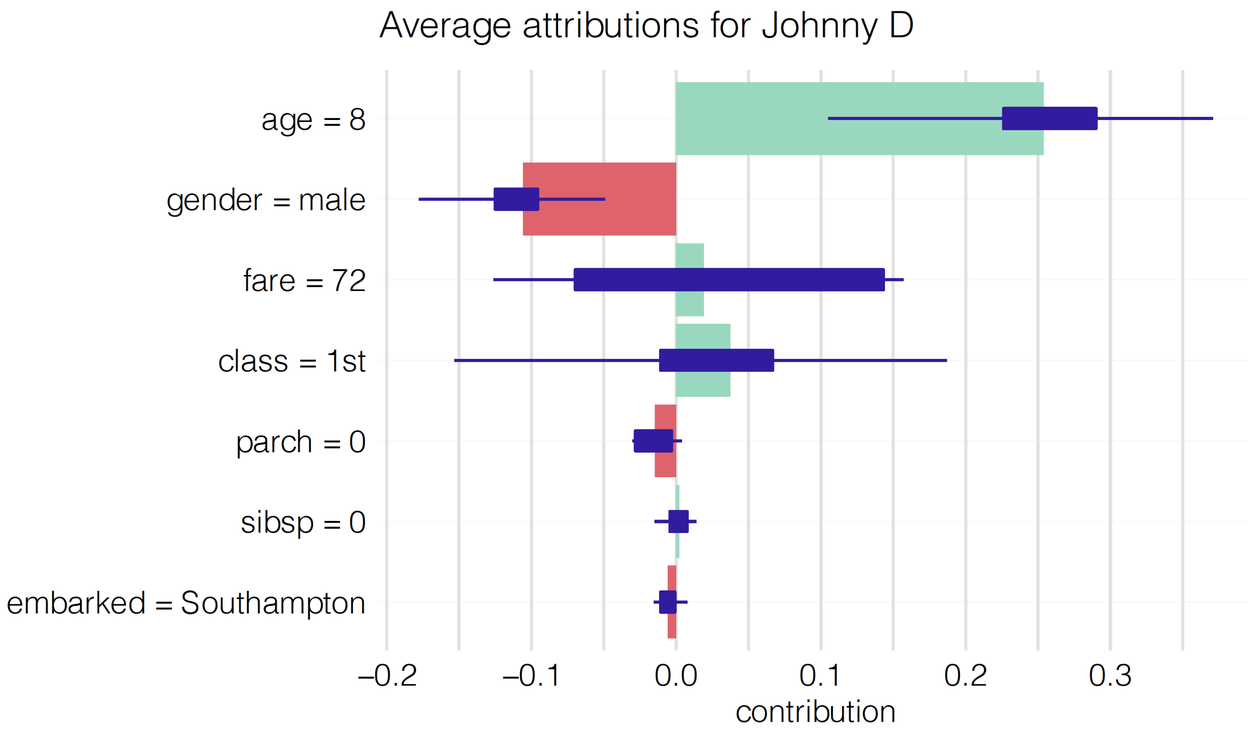


SHAP yöntemi

Değişken sıralamasının etkisini ortadan kaldırmak için, değişken katkılarının ortalama değeri kullanılabilir.

Sırasıyla kırmızı ve yeşil çubuklar, negatif ve pozitif ortalamaları gösterir.

Grafikte en önemli değişkenlerin ***age***, ***class*** ve ***gender*** olduğu görülmektedir.



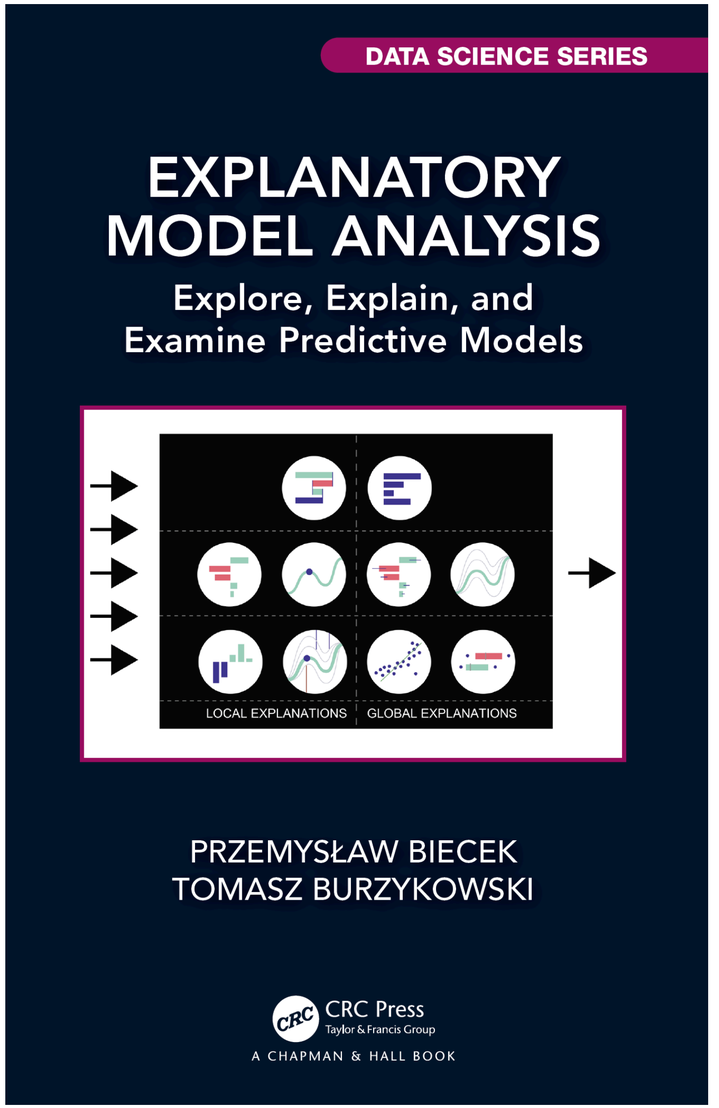
Artı ve eksileri

- Oyun Teorisi'nden devşirilen güçlü bir matematiksel temele sahiptir.

- Model toplamsal değilse (etkileşim etkileri içeriyorsa) yanıltı olabilir.
- Hesaplama maliyeti oldukça yüksektir.

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