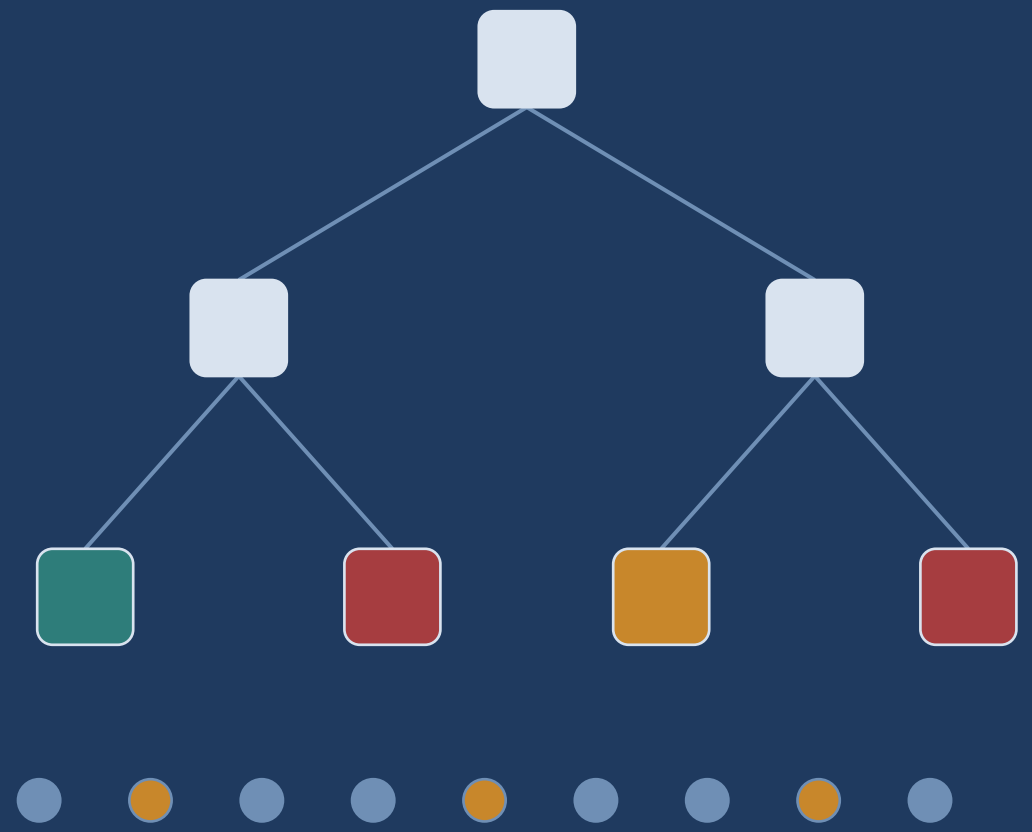
EKONOMİ VE FİNANSTA MAKİNE ÖĞRENMESİ

HAFTA 6

Sorularla Bölmek

Ağaçlar ve Topluluklar

Karar ağacı, rassal orman, boosting ve SHAP ile yorumlanabilirlik



Prof. Dr. Mehmet Songur

Dicle Üniversitesi · İİBF · İktisat Bölümü

Geçen haftadan

Olasılıktan karara



Lojistik regresyon

Doğrusal skor → S eğrisi → olasılık



Karışıklık matrisi

Doğruluk, duyarlılık, kesinlik, özgüllük



ROC ve AUC

Eşikten bağımsız sıralama başarısı



Eşik ve maliyet

Dengesiz veride doğruluk yanıltır

Lojistik regresyonun varsayımı: her değişkenin etkisi tek bir sayıdır, her düzeyde aynıdır ve diğer değişkenlerden bağımsızdır. Gerçek böyle değilse?

Bugün ne konuşacağız?

1

Karar ağacı

Veriyi evet/hayır sorularıyla bölmek

2

Rassal orman

Çok sayıda ağacın ortalaması

3

Boosting

Küçük ağaçlarla hataları art arda düzeltmek

4

SHAP

Kara kutunun her kararını açıklamak

Kredi uzmanı nasıl karar verir?

Bir formülle mi, sırayla sorular sorarak mı?

Doğrusal skor

skor = b_0 + b_1·borç + b_2·gecikme + …

Her değişkene tek bir ağırlık. Borcun etkisi her düzeyde aynı; gecikmenin etkisi borçtan bağımsız.

Soru dizisi

- Borcu gelirine göre çok mu yüksek?
- Evetse: işinde yeni mi?
- Evetse: daha önce gecikmesi var mı?

Soru dizisi eşik koyar ve "şuysa, şuna bak" der. Tek bir katsayı bunu ifade edemez.

Tek ağaç anlaşılır ama oynak; çok ağaç güçlü ama kara kutu; SHAP o kutuyu açar.

Bu haftanın tek cümlesi

1

Karar ağacı

Veriyi sorularla bölmek

Veriden bir soru dizisi kurmak

Kredi uzmanının sorularının otomatik hâli

1

Tüm başvuranlarla başla

3.000 kişi, temerrüt oranı %17

2

En iyi soruyu bul

Her değişkeni ve her eşiği dene; temerrüt edenleri ve ödeyenleri en iyi ayıran soruyu seç

3

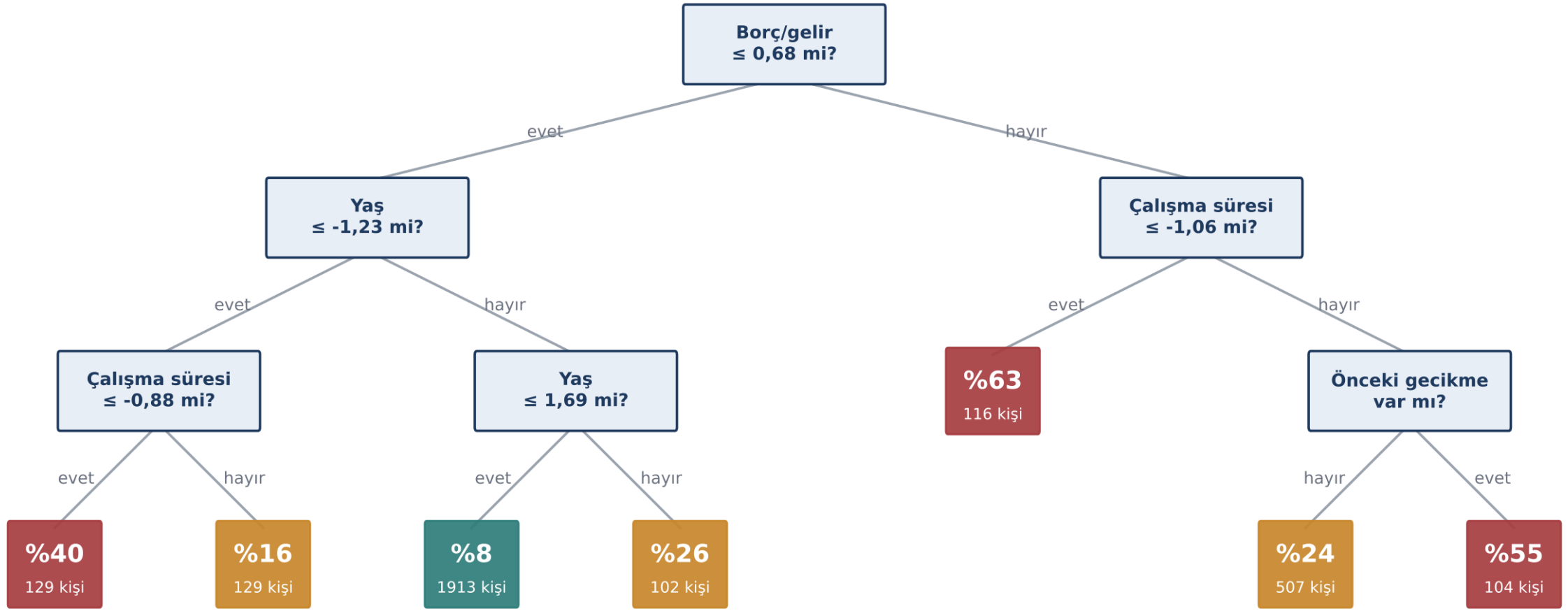
İki grupta tekrarla

Oluşan her grupta aynı işlemi yap; bir sınıra kadar devam et

“En iyi soru”: grupları olabildiğince saf yapan soru (hep temerrüt ya da hep ödeme). Saflığın ölçüsü: safsızlık.

Veriden kurulan bir ağaç

Derinlik 3 · her yaprakta en az 100 kişi · yapraklarda temerrüt oranı



Yapraklardaki sayı: o gruptaki başvuranların temerrüt oranı (eğitim verisi)

Ağacın kendiliğinden bulduğu üç şey

Üçü de simülasyondaki gerçek kurguda var



Eşik

İlk soru: borç/gelir ≤ 0,68 mi? Gerçek kurguda risk 0,7'de sıçrıyor.



Etkileşim

"Gecikme var mı?" yalnızca borcu yüksek olanlara soruluyor: gecikmenin önemi borca bağlı.

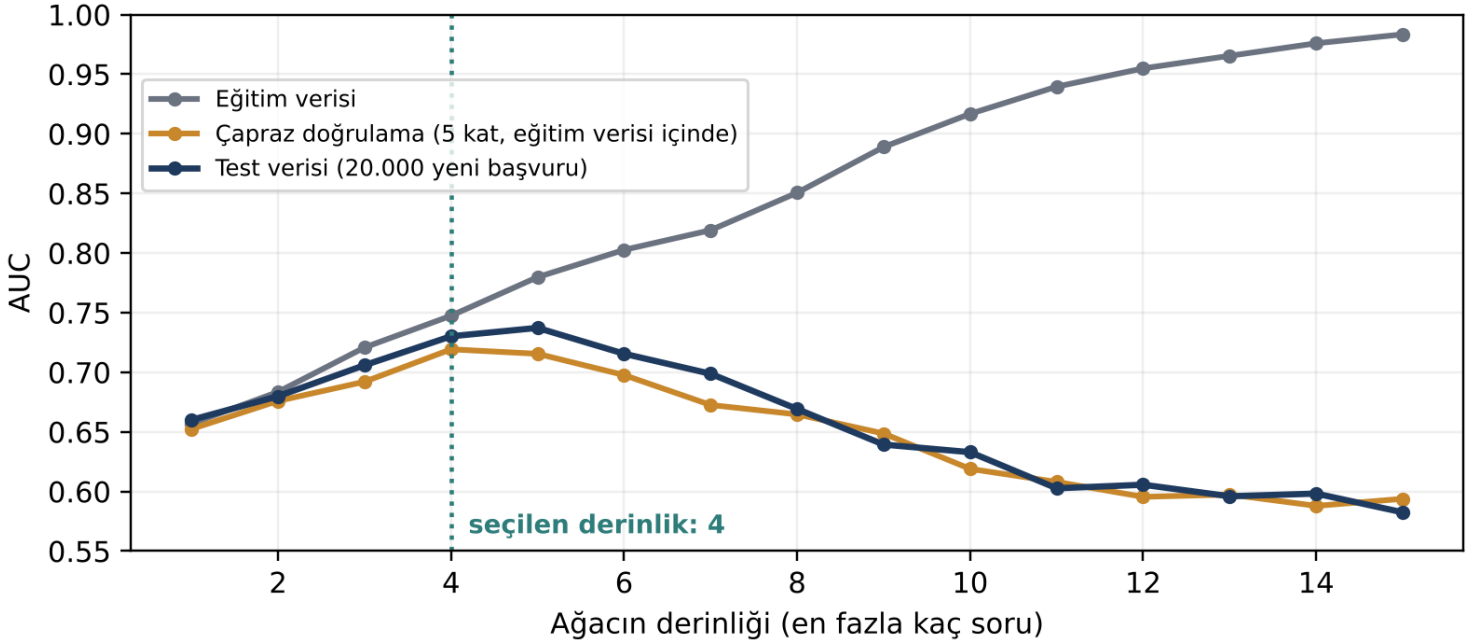


U biçimi

Yaş iki kez soruluyor: çok gençler (%40) ve çok yaşlılar (%26) orta yaşlılardan (%8) riskli.

Ne zaman durmalı? Derinlik ve aşırı uyum

Hafta 3: eğitim verisini ezberleyen model yeni veride başarısız olur



Derinlik 15

Eğitim AUC 0,983
Test AUC 0,583

Derinlik 4 (çapraz doğrulama)

Test AUC 0,730. Derinlik, Hafta 4'teki λ gibi bir karmaşıklık ayarı.

Tek ağaç oynaktır

Aynı büyüklükte 5 farklı eğitim örneklemi, derinliği 4 olan 5 ağaç

Örneklem	İlk soru	İkinci düzeydeki sorular	Test AUC
1	Borç/gelir ≤ 0,70	Yaş · Yaş	0,719
2	Borç/gelir ≤ 0,70	Yaş · Çalışma süresi	0,732
3	Borç/gelir ≤ 0,70	Yaş · Yaş	0,743
4	Borç/gelir ≤ 0,68	Yaş · Önceki gecikme	0,712
5	Borç/gelir ≤ 0,70	Yaş · Çalışma süresi	0,716

İlk soru sağlam, alt sorular değişiyor; AUC 0,712 ile 0,743 arasında. Veri biraz değişince kurallar da değişiyor: yüksek varyans.

Tek ağaç: anlaşılır, esnek ama oynak.

Sığ ağaç fazla basit (yanlılık), derin ağaç ezberler ve örneklemden örnekleme değişir (varyans).

2

Rassal orman

Çok ağacın ortalaması

Kalabalığın bilgeliği

Oynak tahminlerin ortalaması, tek bir tahminden daha az oynaktır



Yeniden örnekle

Eğitim verisinden yerine koyarak 3.000’er gözlem çek (bootstrap)



Her örnekleme bir ağaç

Her soruda rastgele seçilmiş 2 değişken arasından seç



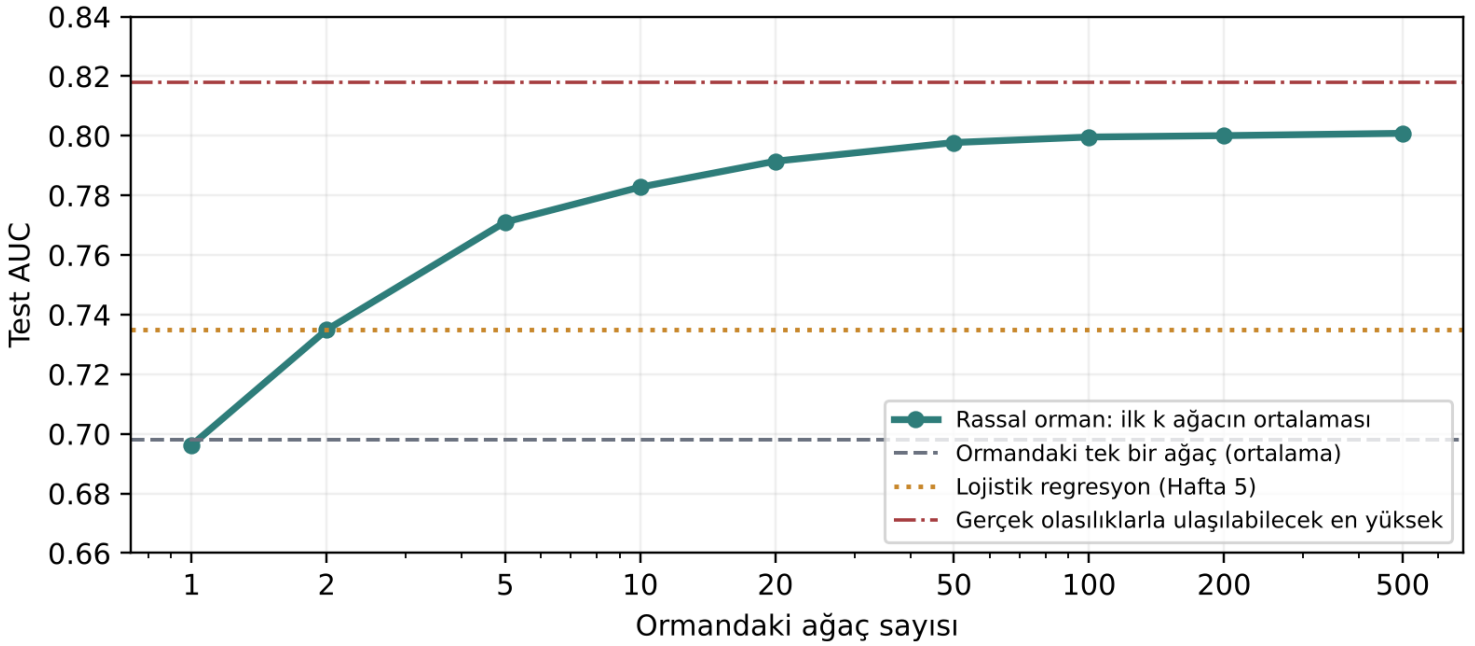
Ortalamasını al

500 ağacın verdiği olasılıkların ortalaması

Neden rastgele değişkenler? Yoksa bütün ağaçlar ilk soruyu borca sorar ve birbirine benzer. Benzer ağaçların ortalaması pek bir şey kazandırmaz.

Ağaç sayısı arttıkça

Test AUC (20.000 yeni başvuru)



Model	AUC
Tek bir ağaç (ortalama)	0,698
Lojistik regresyon	0,735
Bagging (tüm değişkenler)	0,793
Rassal orman	**0,801**
Ulaşılabilecek en yüksek	0,818

Ağaç sayısı aşırı uyum yaratmaz; bir noktadan sonra eğri düzleşir.

Ormanın bedeli: yorumlanabilirlik



Tek ağaç

Tahtaya çizilir, bir başvuranın yolu takip edilir: “borç yüksek, işinde yeni, %63.”



500 ağacın ortalaması

Kimse okuyamaz. “Bu başvuru neden reddedildi?” sorusunun cevabı artık tek bir yol değil.

Bu sorunu 4. bölümde çözeceğiz.

3

Boosting

Hataları art arda düzeltmek

Boosting: küçük ağaçlar, sırayla

Her yeni ağacın görevi, o ana kadarki modelin hatalarını düzeltmek

1

Herkese aynı başlangıç tahmini

2

Küçük bir ağaç kur: en çok yanılınan başvuranlara odaklan

3

Tahminini küçük bir katsayıyla (öğrenme oranı) ekle

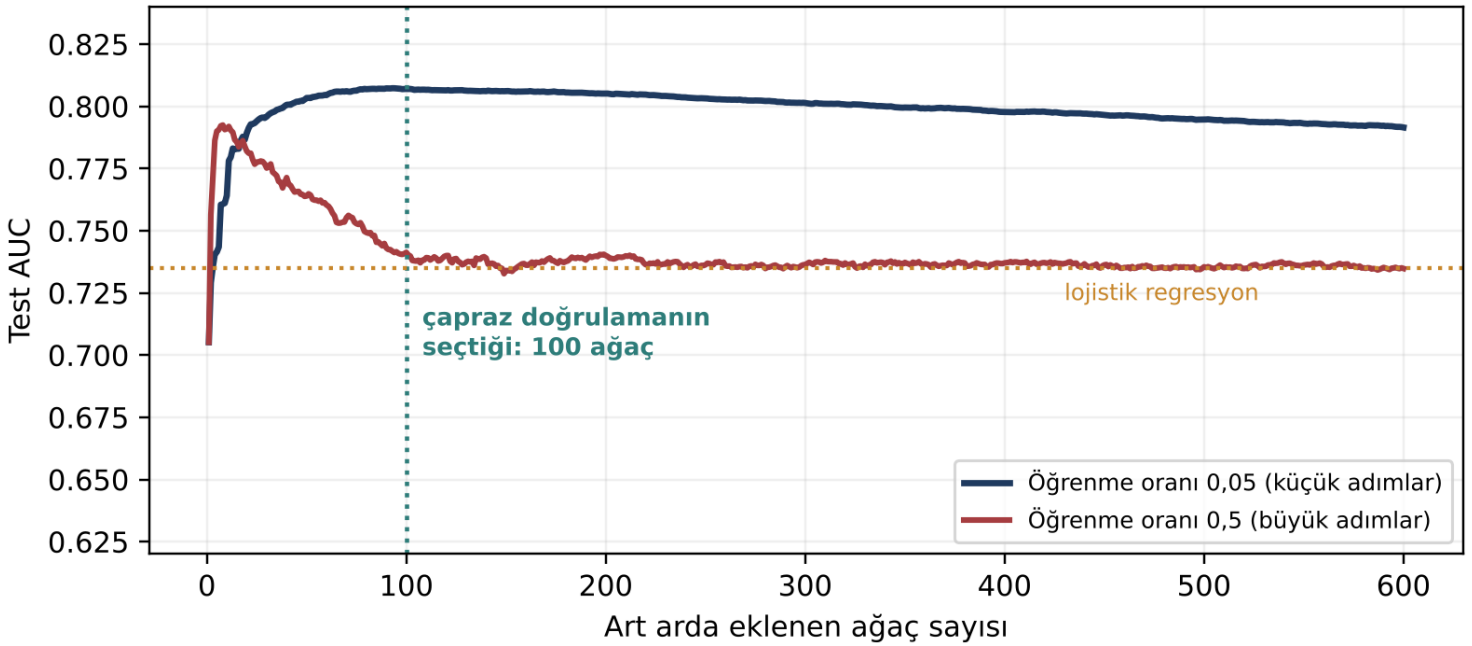
4

Yeni hatalara bak, tekrarla

Rassal orman
Bağımsız ağaçlar, aynı anda, ortalama. Birbirinden habersiz yüz uzmanın oylaması. Varyansı azaltır.

Boosting
Ağaçlar sırayla, her biri öncekinin hatasına bakar. Bir düzeltme zinciri. Yanlılığı azaltır.

İki ayar: ağaç sayısı ve öğrenme oranı



Büyük adımlar (0,5)

8 ağaçta zirve (0,792), 600 ağaçta 0,735: ezber.

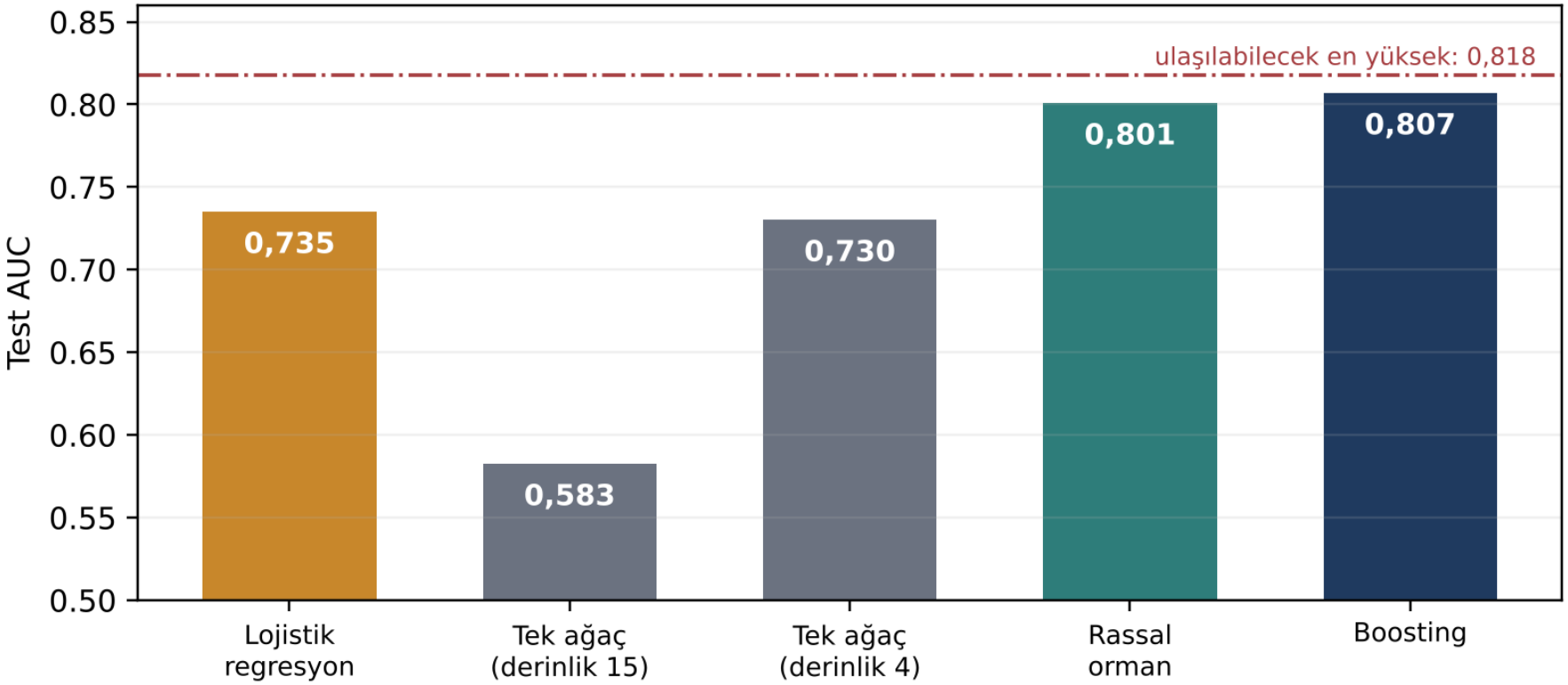
Küçük adımlar (0,05)

Yavaş ama daha yükseğe. Çapraz doğrulama 100 ağaç seçti: test AUC 0,807.

Ormandan farklı olarak boosting'de ağaç sayısı aşırı uyuma yol açar.

Beş model, aynı test verisi

20.000 yeni başvuru



Orman mı, boosting mi?

Rassal orman

- Varsayılan ayarlarla bile çoğu zaman iyi
- Ağaç sayısı zararsız
- Bu örnekte AUC 0,801

Boosting

- Çoğu zaman biraz daha iyi; daha çok ayar ister
- Yanlış ayarla kolayca aşırı uyum
- Bu örnekte AUC 0,807

XGBoost ve LightGBM: boosting'in hızlı ve yaygın uygulamaları. 11. haftada XGBoost kullanacağız.

4

SHAP

Kara kutuyu açmak

“Bu başvuru neden reddedildi?”

Model %62 diyor ama nedenini söylemiyor; katsayı yok



Gerekçe sunmak

Başvurana ve düzenleyiciye



Modeli denetlemek

Mantıklı bir şey mi öğrenmiş?



Veriden öğrenmek

Hangi ilişkileri yakalamış?

Fikir oyun teorisinden: katkıyı adil paylaştırmak

Lloyd Shapley, 2012 Nobel İktisat Ödülü

Shapley'nin sorusu
Oyuncular birlikte bir kazanç elde ettiyse, kazancı katkılarına göre nasıl adil paylaştırırız?

SHAP'in uyarlaması
Oyuncular = değişkenler. Kazanç = bu başvuranın tahmininin ortalama başvurandan farkı.

Bilinen değişkenler (örnek sayılar)	Ortalamadan fark (log-odds)
Hiçbiri	0
Yalnızca borç/gelir	+1,2
Yalnızca önceki gecikme	+0,4
İkisi birden	+2,0

Shapley değeri: her sıradaki katkının ortalaması

Önce borç, sonra gecikme
borç: +1,2
gecikme: 2,0 − 1,2 = +0,8

Önce gecikme, sonra borç
gecikme: +0,4
borç: 2,0 − 0,4 = +1,6

borç = (1,2 + 1,6) / 2 = +1,4 **gecikme = (0,8 + 0,4) / 2 = +0,6**

Toplam +2,0: tahminin ortalamadan farkının tamamı

SHAP’in temel özelliği: toplama

Tahmin = taban (ortalama başvuran) + değişkenlerin SHAP katkıları

−1,93

Taban: ortalama başvuranın log-odds’u

%13

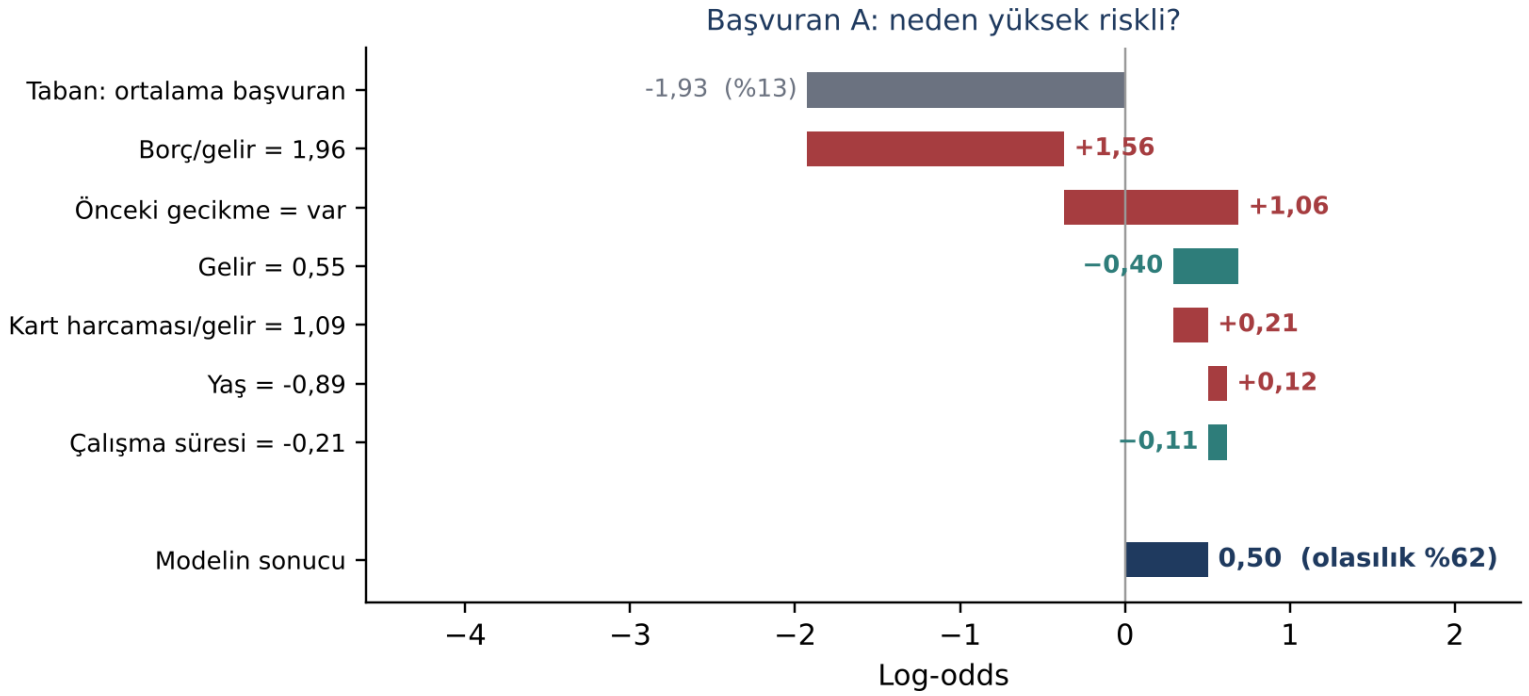
Aynı taban, olasılık olarak

10^{-14}

2.000 başvuruda “taban + katkılar” ile model arasındaki en büyük fark

Başvuran A: neden yüksek riskli?

Kırmızı: riski artıran · yeşil: azaltan (log-odds)

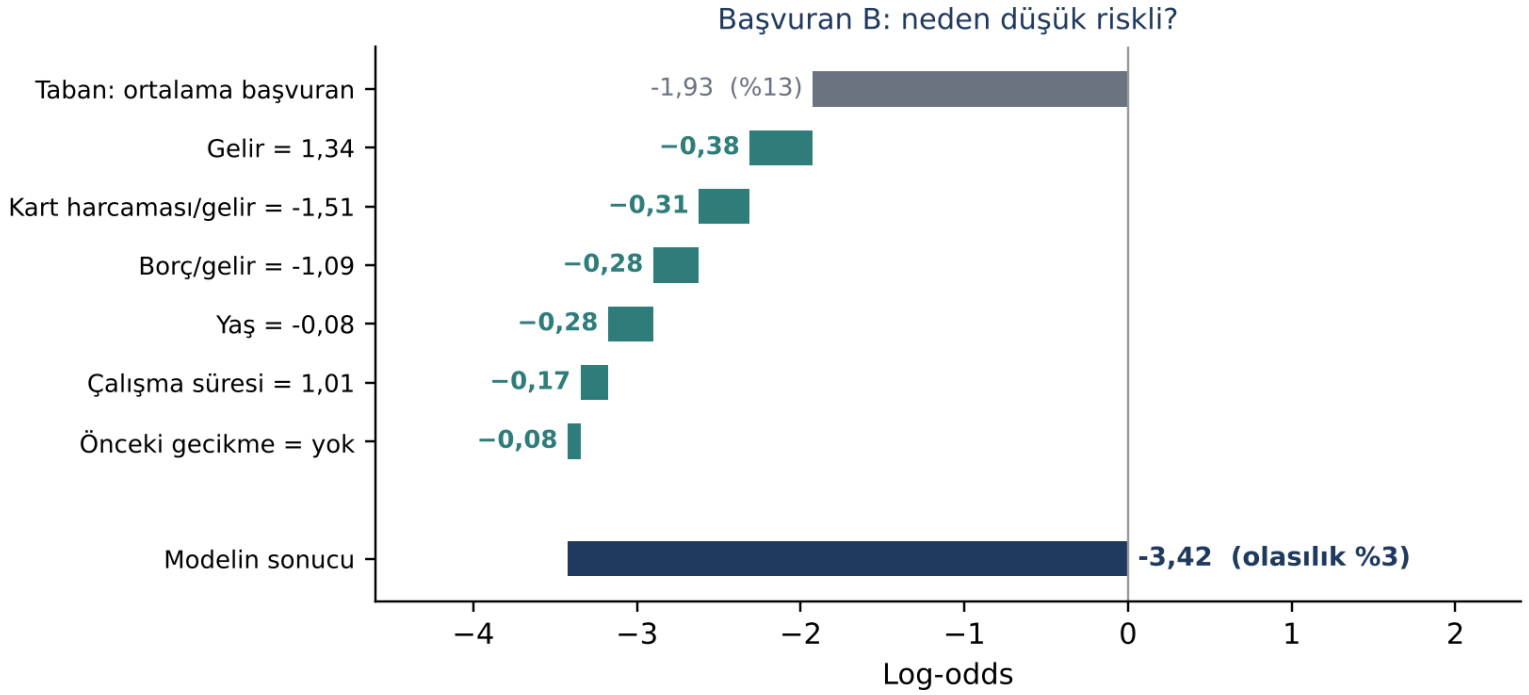


Kredi uzmanının cümlesiyle

"Borcunuz gelirinize göre çok yüksek (+1,56) ve geçmişte gecikmeniz var (+1,06). Geliriniz bunu dengelemeye yetmiyor (−0,40)."

Başvuran B: neden düşük riskli?

Bütün katkılar riski azaltıyor: olasılık %3



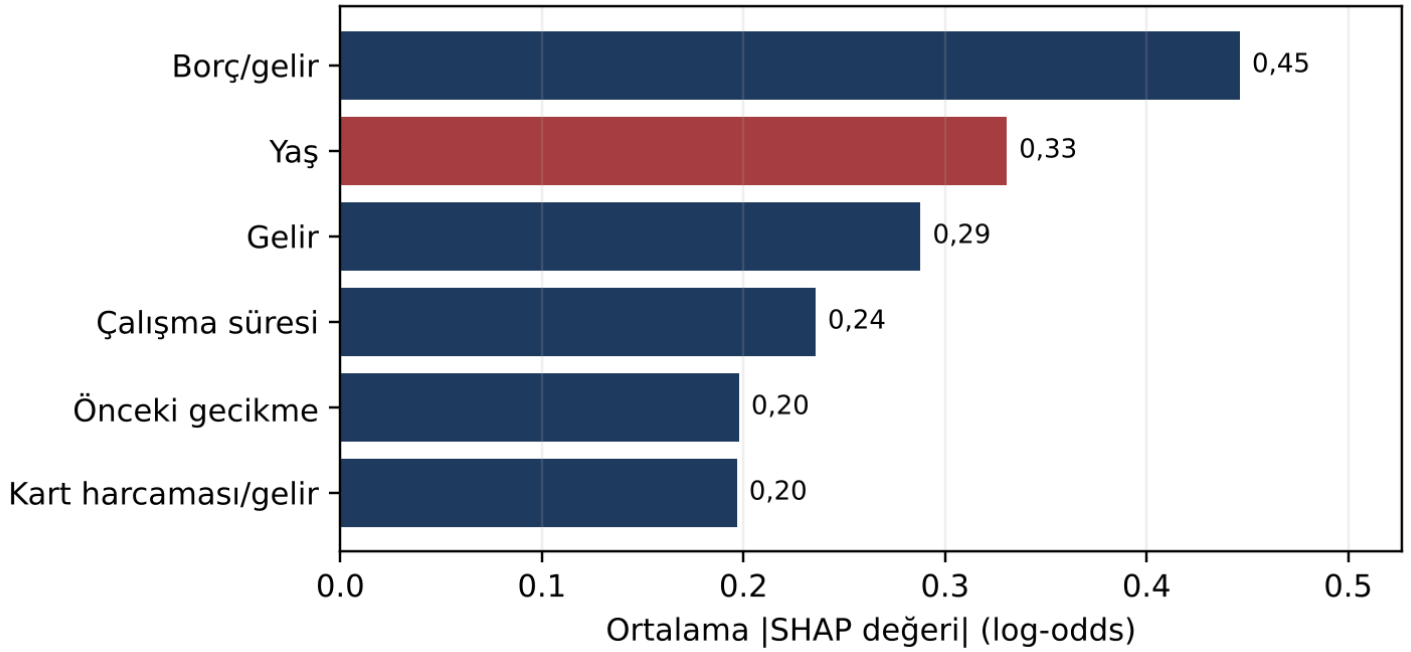
Düşünün

A'nın borç/gelir oranı 1,96'dan 0,5'e inseydi, borcun katkısı yine +1,56 mi olurdu?

SHAP katkısı bir katsayı değildir: aynı değişkenin katkısı başvurandan başvurana değişir.

Bütün model: hangi değişken ne kadar önemli?

2.000 başvuruda ortalama |SHAP katkısı|



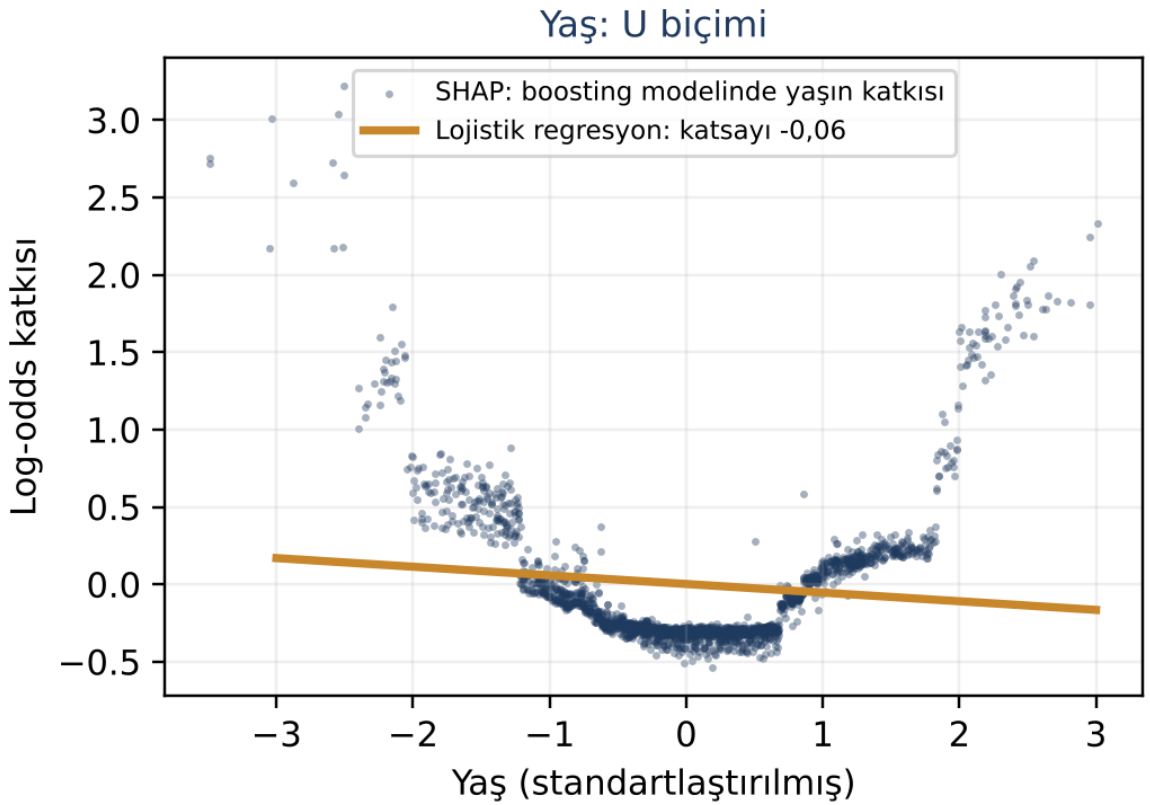
Yaş ikinci sırada (0,33)

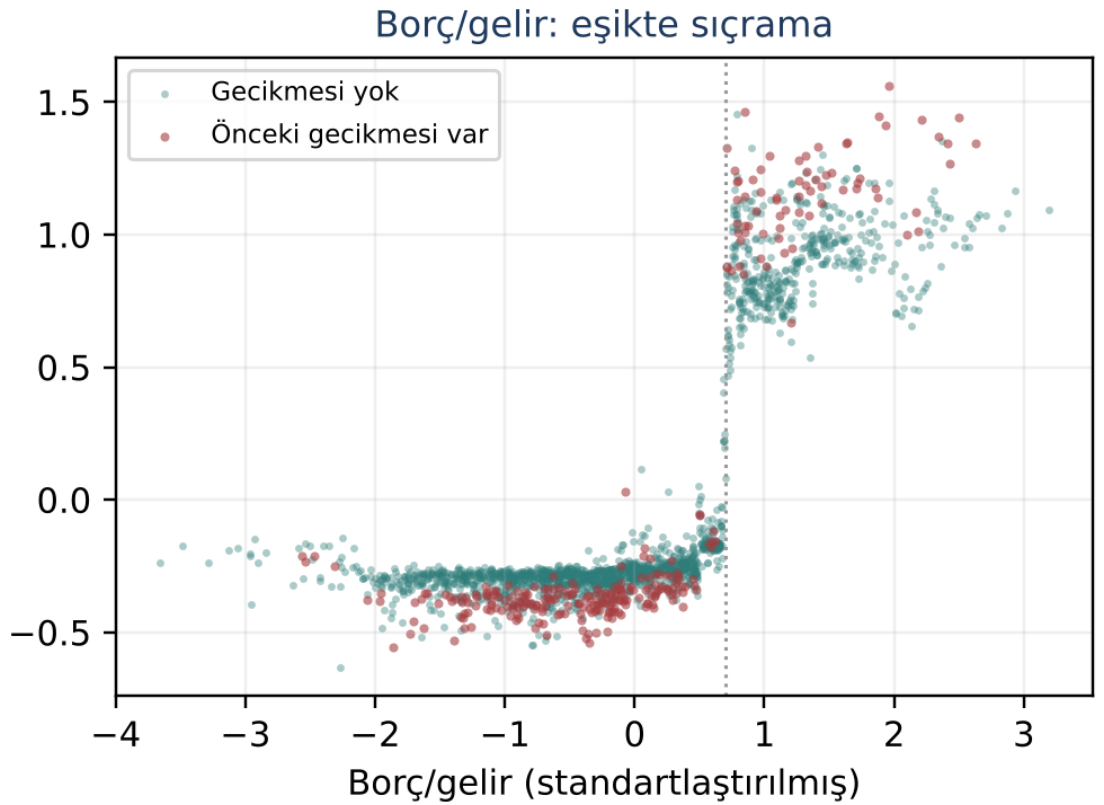
Hafta 5'in lojistik regresyonu yaşa neredeyse sıfır katsayı veriyor: −0,06.

Hangisi doğru?

Bağımlılık grafikleri

Yatay: değişkenin değeri · dikey: o değişkenin SHAP katkısı · her nokta bir başvuran

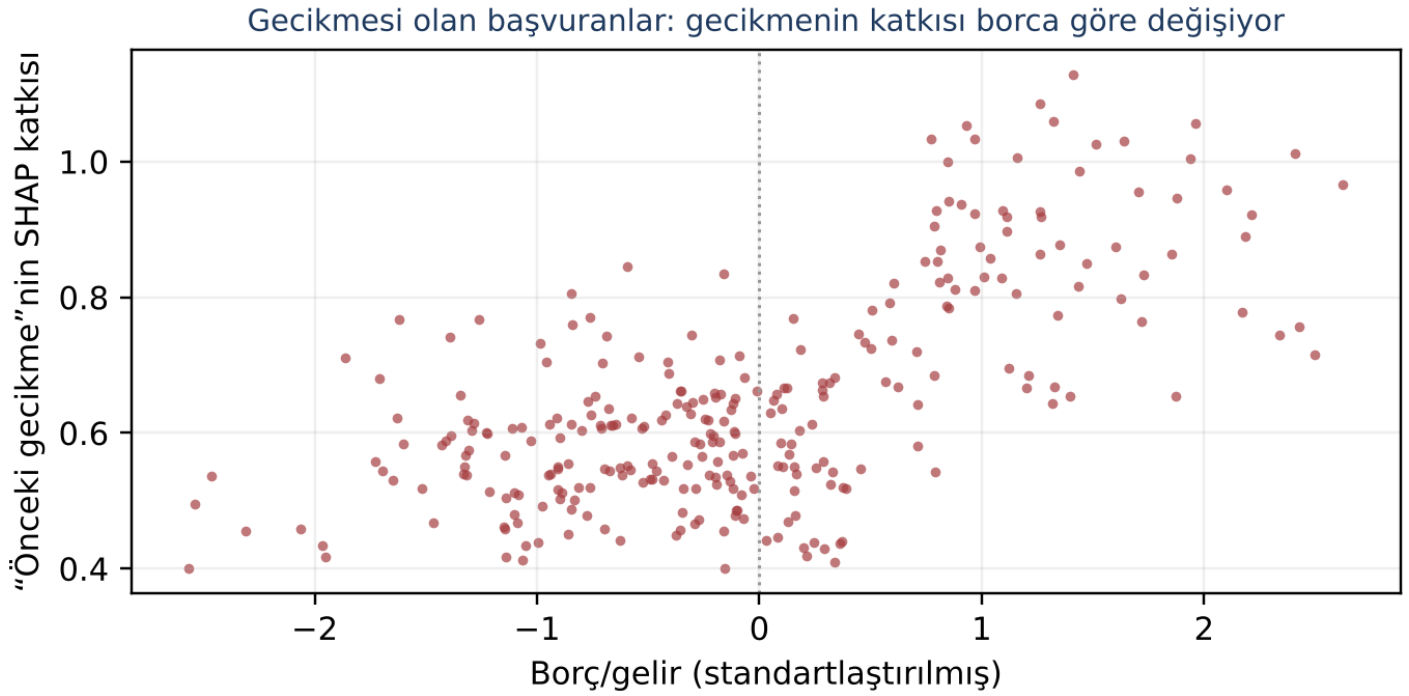




Katsayı sıfıra yakın diye etki yok değildir: doğrusal model U biçimini düz bir çizgiye indiriyor.

Etkileşim: gecikmenin önemi borca bağlı

Önceki gecikmesi olan başvuranlar



0,58

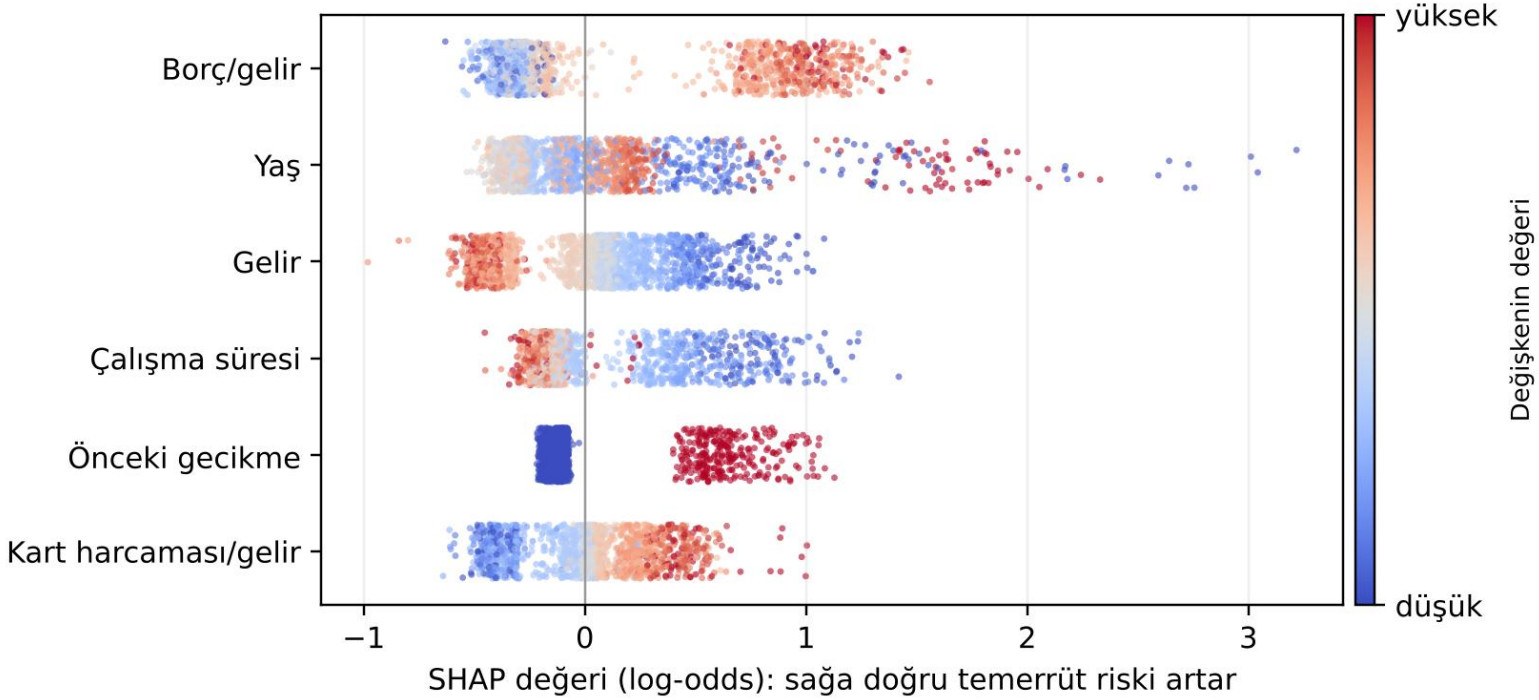
Borcu ortalamanın altında olanlarda gecikmenin katkısı

0,75

Borcu ortalamanın üzerinde olanlarda

Özet grafik: önem ve yön bir arada

Renk: değişkenin değeri (kırmızı yüksek, mavi düşük)



Gelir
Kırmızılar solda: gelir arttıkça risk düşüyor.

Yaş
İki renk de sağ uçta: U biçiminin bu grafikteki görüntüsü.

SHAP neyi söylemez?



Nedensel etki değil

“Borcunu azaltırsa riski şu kadar düşer” demez. Modelin tahminini açıklar, dünyayı değil.



İlişkili değişkenler paylaşır

Yaş ve çalışma süresi ilişkili; katkı ikisi arasında bölünebilir. Düşük SHAP ≠ önemsiz bilgi.



Modeli açıklar, veriyi değil

Model yanlış bir şey öğrendiyse SHAP o yanlışı sadakatle açıklar.

SHAP modeli açıklar, dünyayı değil.

Tahmin = taban + katkılar. Bir SHAP grafiği bir hipotez kaynağıdır, nedensel bir sonuç değil.

5

Uyarılar ve alıştırma

İktisatçı için

İktisatçı için üç uyarı



Esnek model, çok veri ister

3.000 gözlemde iyi çalıştı. 120 aylık bir makro seride aynı esneklik kolayca ezbere döner.



Ağaçlar dışarıya taşamaz

Görülmemiş bir faiz ya da enflasyon düzeyinde en yakın yaprağın değerini verir; eğilimi sürdürmez.



Öngörü ≠ açıklama

Yüksek AUC, ilişkiyi nedensel olarak öğrendiğimiz anlamına gelmez.

Alıştırma: doğru mu, yanlış mı, eksik mi?

Gerekçesiyle cevap verin

1. A’da borcun SHAP katkısı +1,56; borcunu azaltırsa olasılığı bu kadar düşer.
2. A’nın SHAP katkıları ile tabanın toplamı modelin log-odds tahminidir.
3. Lojistik katsayısı sıfıra yakın; yaşın temerrütle ilişkisi yok.
4. Çalışma süresinin SHAP değeri küçükse, bu bilginin ilgisi yoktur.
5. Boosting’de ağaç sayısı arttıkça test başarısı hep artar.
6. Rassal orman, enflasyon tarihsel zirvenin iki katına çıkınca da güvenilirdir.

Cevaplar

SHAP modeli açıklar; dünyayı değil

1	A'da borcun SHAP katkısı +1,56; borcunu azaltırsa olasılığı bu kadar düşer.	**Yanlış: nedensel değil**
2	A'nın SHAP katkıları ile tabanın toplamı modelin log-odds tahminidir.	**Doğru: toplama**
3	Lojistik katsayısı sıfıra yakın; yaşın temerrütle ilişkisi yok.	**Yanlış: U biçimi**
4	Çalışma süresinin SHAP değeri küçükse, bu bilginin ilgisi yoktur.	**Eksik: yaşla paylaşım**
5	Boosting'de ağaç sayısı arttıkça test başarısı hep artar.	**Yanlış: aşırı uyum**
6	Rassal orman, enflasyon tarihsel zirvenin iki katına çıkınca da güvenilirdir.	**Yanlış: dışarıya taşamaz**

Bugünün beş mesajı

1. Karar ağacı veriyi sorularla böler; eşikleri, etkileşimleri ve doğrusal olmayan ilişkileri kendiliğinden yakalar.
2. Tek ağaç ya fazla basittir ya da ezberler ve oynaktır; derinlik çapraz doğrulamayla seçilir.
3. Rassal orman çeşitlendirilmiş ağaçların ortalamasıyla varyansı azaltır; ağaç sayısı aşırı uyum yaratmaz.
4. Boosting küçük ağaçları art arda ekleyerek hataları düzeltir; ağaç sayısı ve öğrenme oranı birlikte seçilir.
5. SHAP her tahmini “taban + katkılar” olarak yazar; tek kararı ve bütün modeli açıklar, ama nedensel etki göstermez.