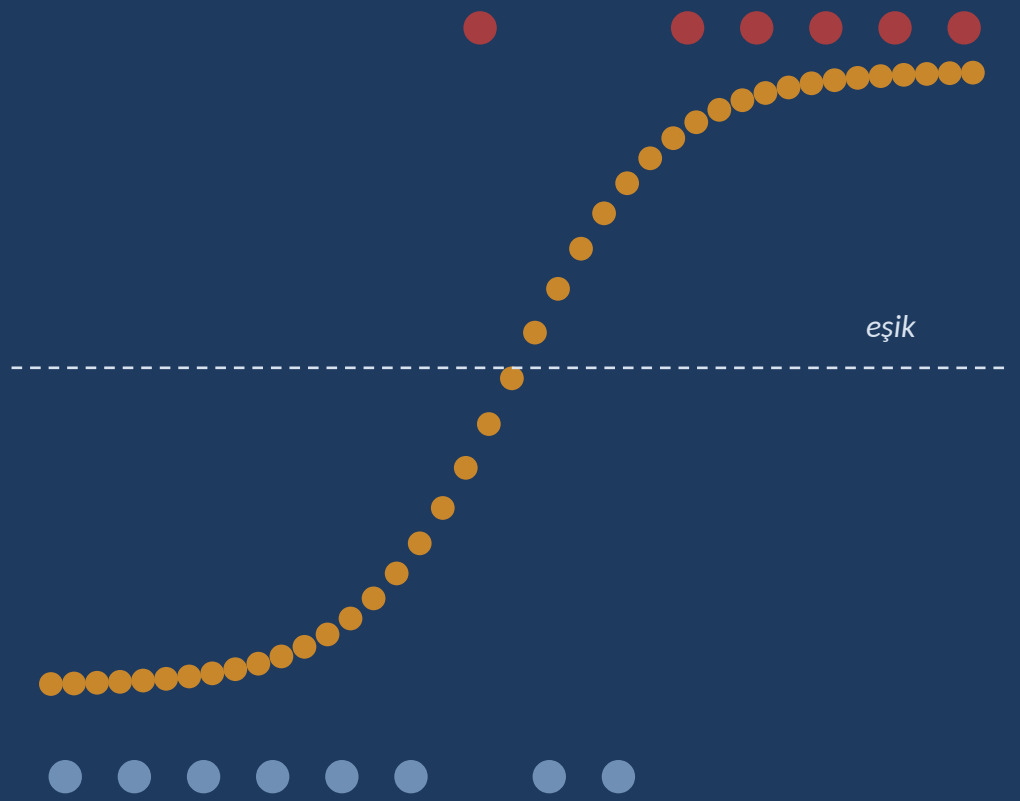
EKONOMİ VE FİNANSTA MAKİNE ÖĞRENMESİ

HAFTA 5

Olasılıktan Karara

Sınıflandırma

Lojistik regresyon, karışıklık matrisi, ROC-AUC, dengesiz veri



Prof. Dr. Mehmet Songur

Dicle Üniversitesi · İİBF · İktisat Bölümü

Geçen haftadan

Çok değişken, az gözlem: katsayılara ceza kesmek

Hata + λ × ceza

Biraz yanlılık, çok daha az varyans

Ridge

Hepsini küçültür, hiçbirini sıfırlamaz

Lasso

Bazılarını tam sıfır yapar: köşe çözümü

λ seçimi

Çapraz doğrulamayla; önce standartlaştır

Hafta 2'de verilen söz: "Sınıflandırma modelinin verdiği olasılığın (0,18) nasıl bir karara çevrileceği 5. haftanın konusudur." Bugün o hafta.

Bugün ne konuşacağız?

1

Lojistik regresyon

Hedef 0 ya da 1 iken olasılık nasıl kestirilir?

2

Eşik ve karışıklık matrisi

Olasılıktan karara; dört kutu, dört ölçü

3

ROC ve AUC

Tüm eşikleri bir arada değerlendirmek

4

Maliyet ve dengesiz veri

Eşiği ne belirler? Olay nadirse ne olur?

%90 doğru bir model

Temerrüt modeli, 20.000 yeni başvuruda test edildi. İyi bir model mi?

%90

Modelin doğru sınıflandırdığı başvurular

%89,3

Herkese "öder" deseydik doğruluk

%14

Temerrüde düşenlerden yakalanan pay

Başvuranların yalnızca %11'i temerrüde düşüyor. Model, hiçbir şey yapmayan kuraldan neredeyse ayırt edilemiyor ve asıl önemli grubu kaçırıyor. Bu sonucu model aynı kalsa bile değiştirebilir miyiz?

Model bir olasılık verir; kararı eşik verir; eşiği hataların maliyeti belirler.

Bu haftanın tek cümlesi

1

Lojistik regresyon

Hedef 0 ya da 1 olduğunda olasılık nasıl kestirilir?

Temerrüdü 0/1 kodlayıp EKK kurarsak?

Ekonometrideki adıyla: doğrusal olasılık modeli

Olasılık 0–1 dışına taşar

%16,5

test başvurusuna negatif temerrüt olasılığı veriyor. “Olasılık −%4” anlamsız.

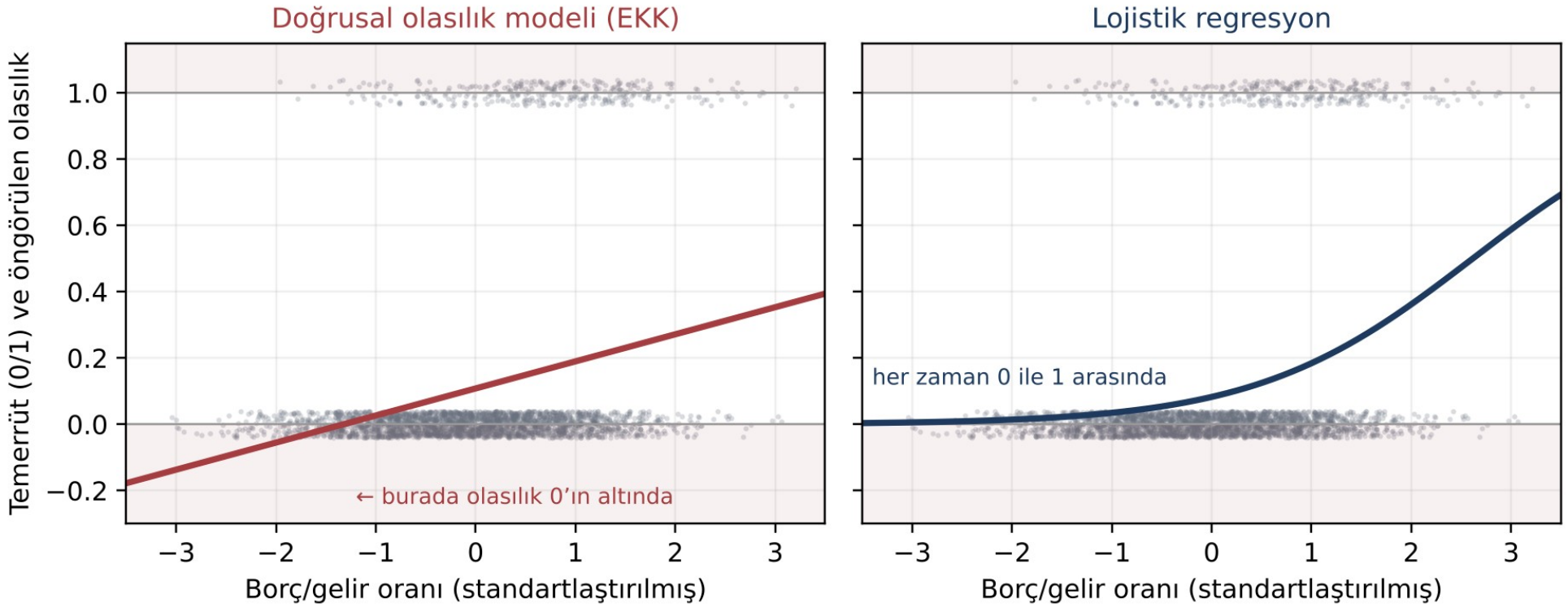
Etki her yerde aynı

Doğru, borçtaki bir artışın olasılığı her düzeyde aynı miktarda artırdığını varsayar. Oysa düşük riskli biri %1’den %2’ye, sınırdaki biri %40’tan %60’a çıkabilir.

İhtiyacımız olan: her zaman 0 ile 1 arasında kalan ve uçlarda yavaşlayan bir olasılık eğrisi.

Doğru mu, S eğrisi mi?

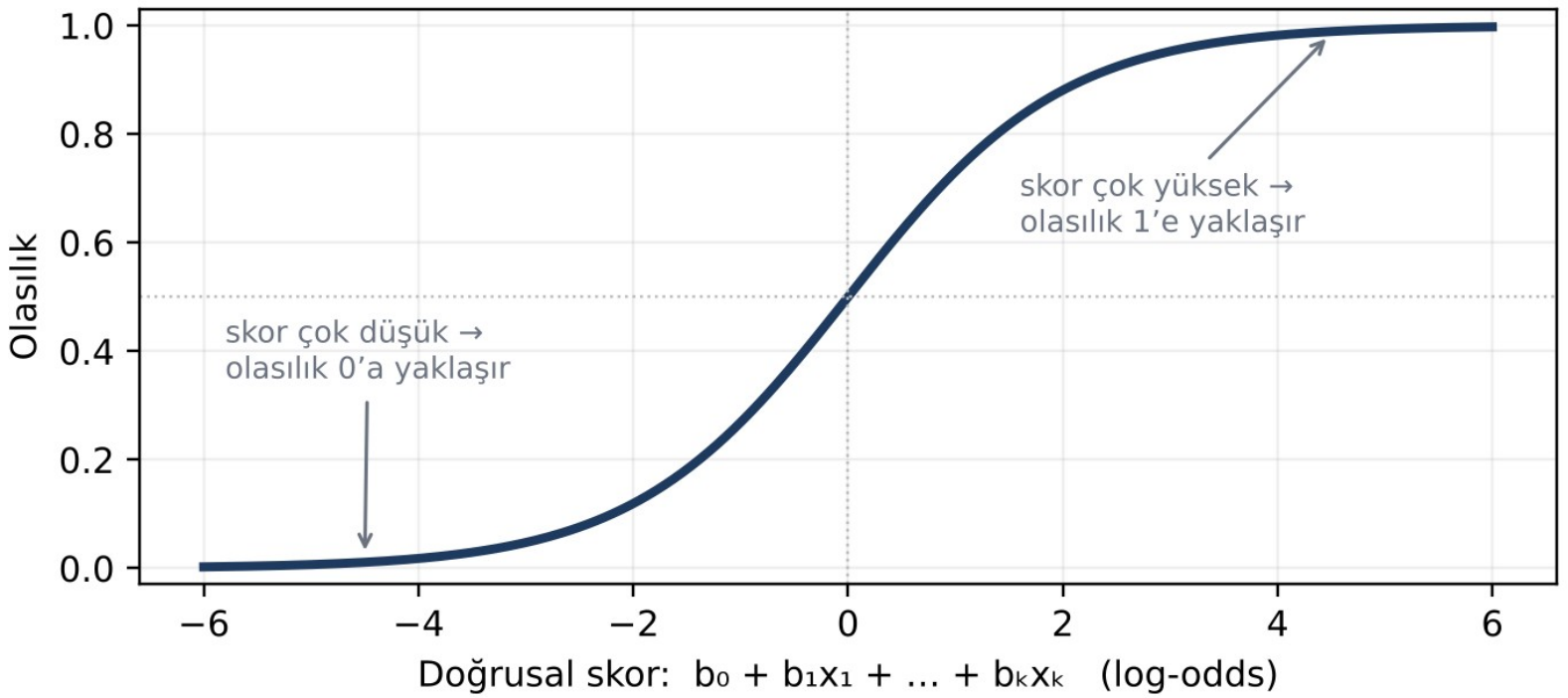
Tek değişken: borç/gelir oranı. Noktalar gerçek sonuçlar (0 ya da 1).



Lojistik regresyon iki adımdır

1 **Skor** $= b_0 + b_1x_1 + \ldots + b_kx_k$

2 **Olasılık** $= 1 / (1 + e^{-\text{skor}})$



Ekonometriden tanıdığınız model

Makine öğrenmesindeki lojistik regresyon = ekonometrideki logit

Log-odds

Skor, “temerrüt / ödeme” oranının logaritmasıdır: ln[p / (1 − p)].

En çok olabilirlik

Katsayılar, gözlenen 0’ları ve 1’leri en olası kılacak biçimde seçilir.

Bir sınıflandırma yöntemi

Adında “regresyon” geçse de amacı bir sınıfın olasılığını kestirip sınıf tahmini yapmaktır.

Katsayılar: odds oranı

e^b: değişken bir birim artınca temerrüt odds'u kaç katına çıkıyor? (3.000 başvuruyla eğitilen model)

× 2,2

Borç/gelir oranında 1 standart sapma artış

× 2,6

Önceki gecikmesi olmak

× 0,55

Gelirde 1 standart sapma artış

Odds oranı olasılık oranı değildir: olasılıktaki değişim kişinin başlangıç riskine bağlıdır. Ayrıca bunlar bir öngörü modelinin katsayıları; bu haftanın konusu olasılıkların ne kadar işe yaradığı.

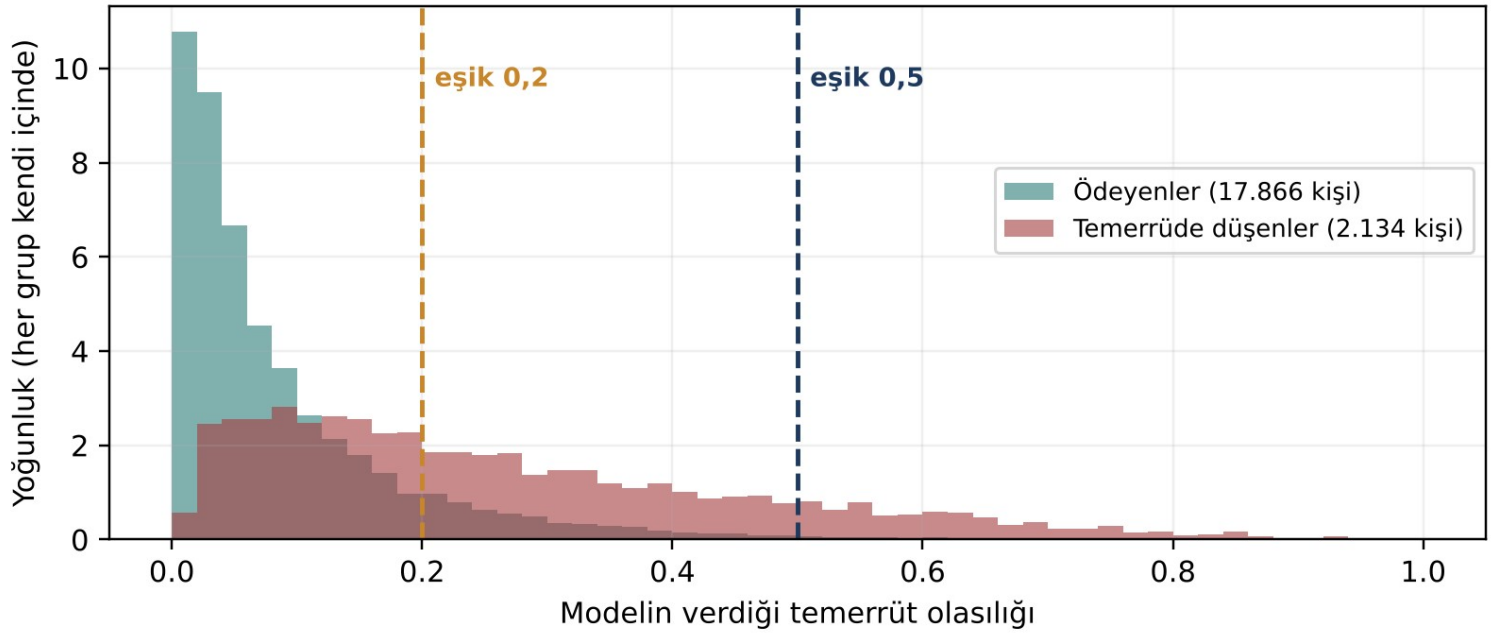
2

Olasılıktan karara

Eşik ve karışıklık matrisi

Eşik: olasılığı karara çevirmek

Temerrüt olasılığı eşiğin üzerindeyse reddet, altındaysa onayla



İki dağılım iç içe: hangi eşik seçilirse seçilsin bazı hatalar kaçınılmaz.

Eşik 0,5

Yazılımların varsayılanı. Temerrüde düşenlerin çoğu bu çizginin solunda: “öder” deniyor.

Eşik 0,2

Çok daha fazla temerrüt yakalanır; karşılığında daha fazla iyi müşteri reddedilir.

Karışıklık matrisi

Bir eşik seçildiğinde her başvuru dört kutudan birine düşer. Pozitif sınıf: temerrüt.

	Model: temerrüt (reddet)	Model: öder (onayla)
Gerçek: temerrüt	**Doğru pozitif (DP)** Riskli başvuru doğru biçimde reddedildi	**Yanlış negatif (YN)** Temerrüde düşecek kişiye kredi verildi
Gerçek: öder	**Yanlış pozitif (YP)** Ödeyecek bir müşteri reddedildi	**Doğru negatif (DN)** İyi müşteri onaylandı

Bankaya maliyeti: YN → geri dönmeyen anapara. YP → kaçırılan faiz geliri. İki hata eşit değil.

Aynı model, iki eşik

20.000 test başvurusu

Eşik 0,5

	Model: temerrüt (reddet)	Model: öder (onayla)
Gerçek: temerrüt	**308** Doğru pozitif (DP)	**1.826** Yanlış negatif (YN)
Gerçek: öder	**168** Yanlış pozitif (YP)	**17.698** Doğru negatif (DN)

Eşik 0,2

	Model: temerrüt (reddet)	Model: öder (onayla)
Gerçek: temerrüt	**1.147** Doğru pozitif (DP)	**987** Yanlış negatif (YN)
Gerçek: öder	**2.122** Yanlış pozitif (YP)	**15.744** Doğru negatif (DN)

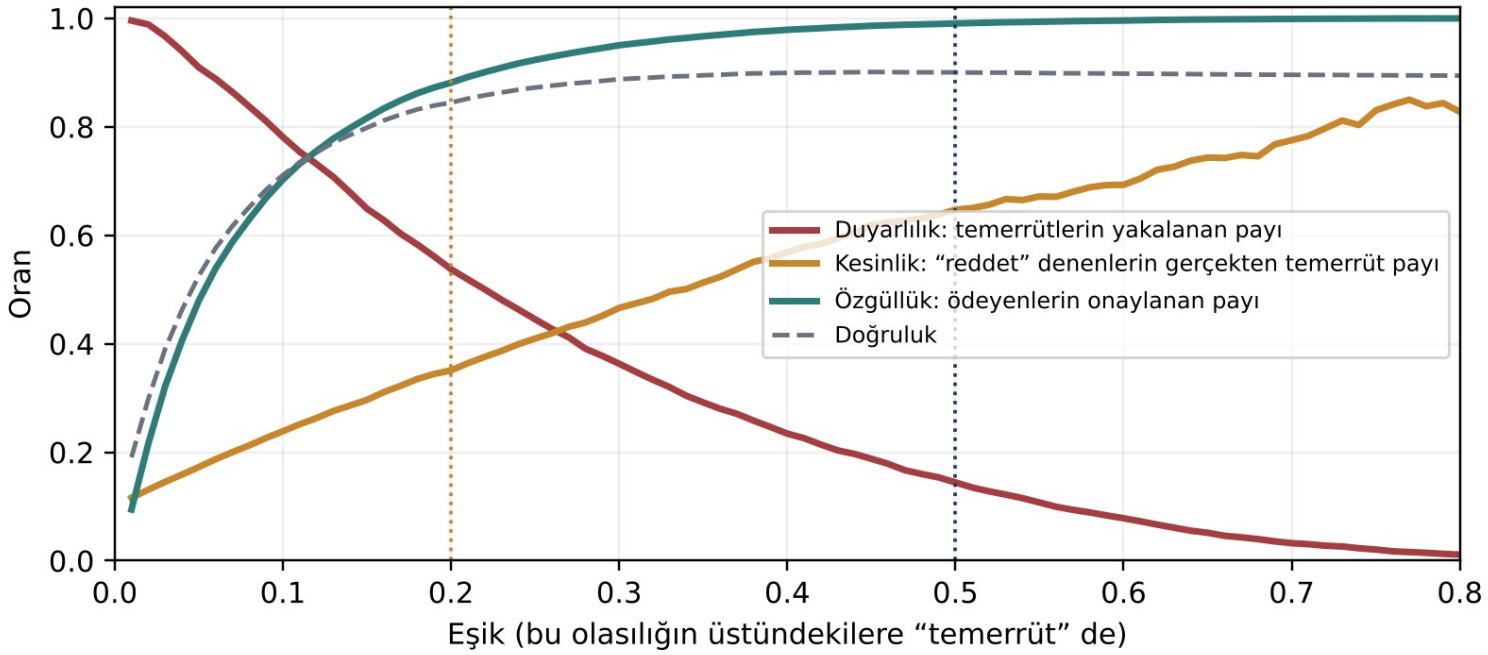
Dört ölçü, dört farklı soru

Ölçü	Formül	Cevapladığı soru	Eşik 0,5	Eşik 0,2
Doğruluk	(DP + DN) / hepsi	Kararların yüzde kaçı doğru?	%90,0	**%84,5**
Duyarlılık	DP / (DP + YN)	Temerrüde düşenlerin yüzde kaçını yakaladık?	%14,4	**%53,7**
Kesinlik	DP / (DP + YP)	“Reddet” dediklerimizin yüzde kaçı gerçekten temerrüde düştü?	%64,7	%35,1
Özgüllük	DN / (DN + YP)	Ödeyenlerin yüzde kaçını onayladık?	%99,1	%88,1

Eşik inince doğruluk düşüyor, yakalanan temerrüt neredeyse dörde katlanıyor. Hangi model daha iyi?

Eşik değiştikçe: ödünleşim

Birini artırmak için eşiği oynattığımızda diğeri düşüyor



Eşik ↑

Duyarlılık düşer; özgüllük ve kesinlik artar.

Doğruluk

0,5 civarında en yüksek: model neredeyse herkese “öder” diyor, çoğunluk da zaten ödüyor.

3

ROC eğrisi ve AUC

Tüm eşikleri tek bir resimde görmek

ROC eğrisi nasıl kurulur?

Her eşik için iki oran hesapla, bir nokta olarak işaretle

Dikey eksen: duyarlılık

Temerrüde düşenlerden yakalanan pay. Yüksek olsun isteriz.

Yatay eksen: yanlış pozitif oranı

Ödeyenlerden yanlışlıkla reddedilen pay = 1 − özgüllük. Düşük olsun isteriz.

Eşik 1

kimse reddedilmez: sol alt köşe (0, 0)

Eşik ↓

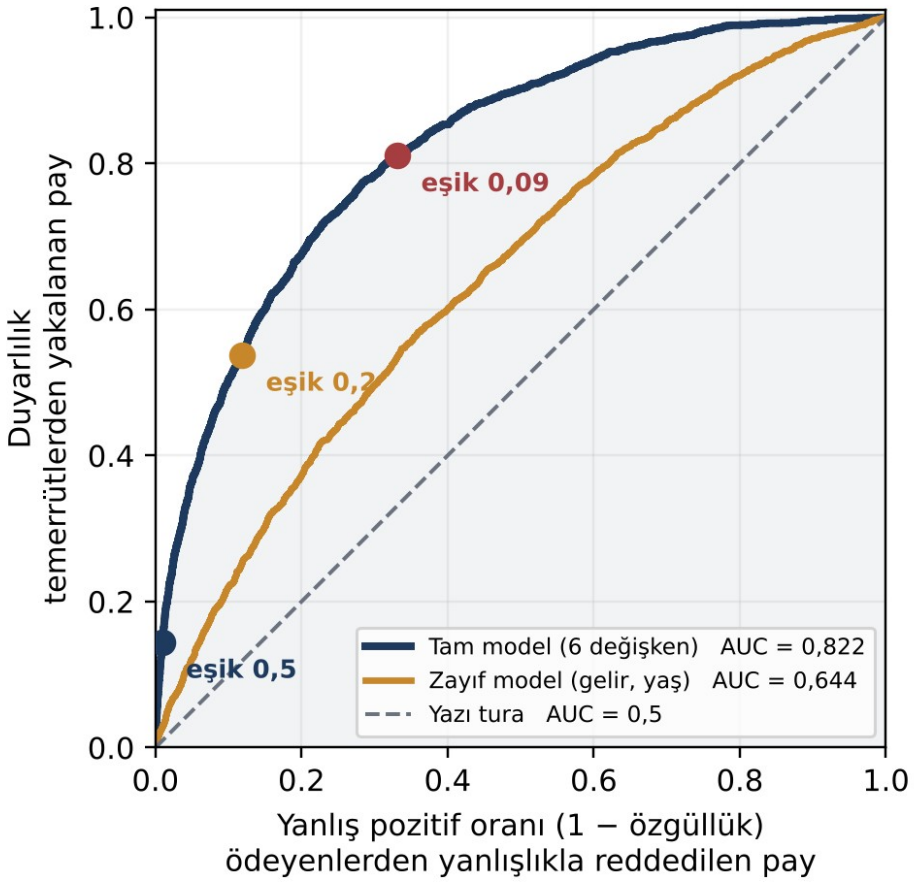
nokta sağ üste doğru ilerler

Eşik 0

herkes reddedilir: sağ üst köşe (1, 1)

ROC eğrileri

Sol üste ne kadar yaslanırsa o kadar iyi



Sol üst köşe

Mükemmel model: tüm temerrütler yakalanır, kimse haksız yere reddedilmez.

Köşegen

Yazı tura: rastgele reddetmek iki oranı eşit artırır.

Tam model her yerde üstte

Hangi eşik seçilirse seçilsin zayıf modelden daha iyi bir ödünleşim sunuyor.

AUC: eğrinin altındaki alan

Mükemmel model 1, yazı tura 0,5

0,822

Tam model

0,644

Zayıf model (gelir, yaş)

0,500

Yazı tura



Olasılık yorumu: rastgele bir temerrüt edenle rastgele bir ödeyeni yan yana koyun. Model temerrüt edene daha yüksek olasılık veriyor mu?

100.000 rastgele çift: model çiftlerin 0,820 kadarında doğru sıralıyor. AUC ile aynı.

AUC’nin sınırları



Eşik seçmez

AUC’si yüksek bir model kötü bir eşikle kullanılırsa sonuç yine kötüdür (eşik 0,5’teki modelimiz gibi).



Olasılıkları ölçmez

Bütün olasılıkları ikiye katlasak sıralama değişmez, AUC de değişmez. Fiyatlamada olasılığın kendisi önemlidir.



1’e ulaşılamaz

Gerçek olasılıkları bilsek bile en yüksek AUC 0,823. Temerrütte her zaman bir şans payı var.

AUC eşikten bağımsızdır; doğruluk, duyarlılık ve kesinlik eşiğe bağlıdır.

AUC: “model riskli olanı ayırt edebiliyor mu?” Karışıklık matrisi: “seçtiğimiz kararın sonuçları ne?”

4

Eşiği maliyet belirler

Hafta 1: “ya da daha iyisi, bankanın beklenen zararı”

İki hatanın bedeli aynı değil

Basit bir varsayım: kaybedilen anapara, kaybedilen faiz gelirinden çok daha büyük

10

Yanlış negatif

Temerrüde düşecek birine kredi vermek

1

Yanlış pozitif

Ödeyecek birini reddetmek

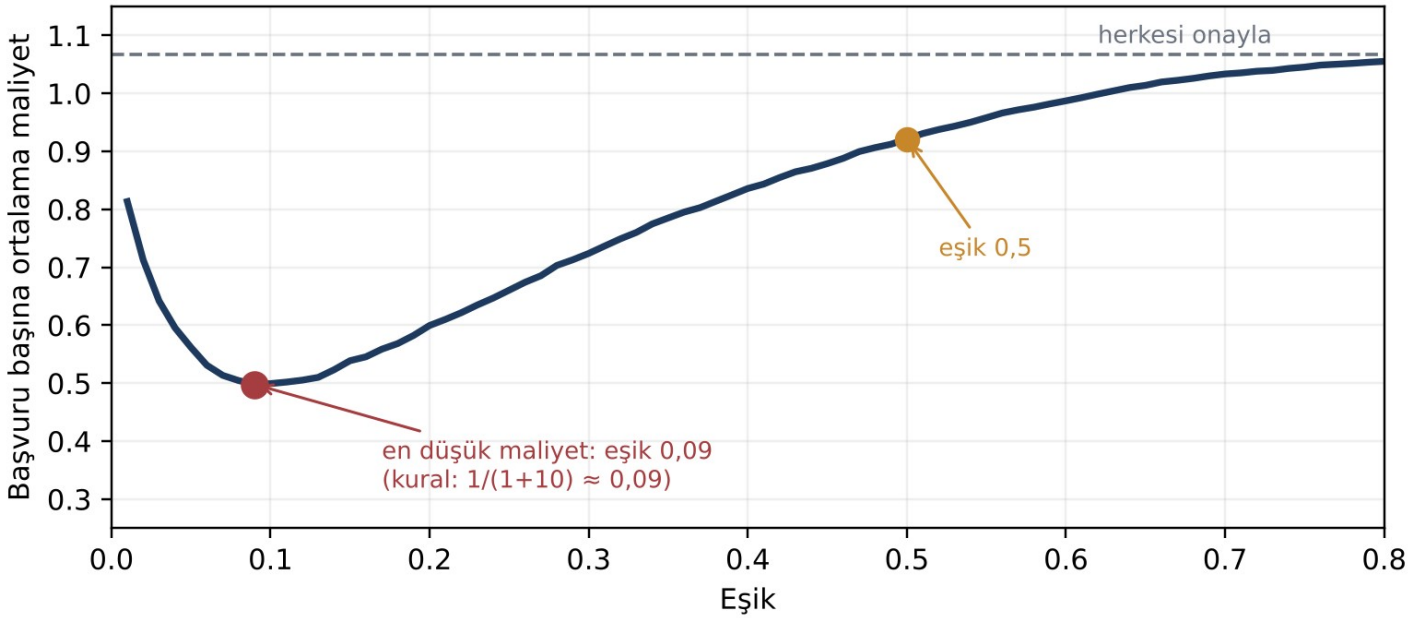
0

Doğru kararlar

DP ve DN

$$\text{maliyet} = (10 \times \text{YN} + 1 \times \text{YP}) \;/\; \text{başvuru sayısı}$$

Her eşik için ortalama maliyet



Karar kuralı	Duyarlılık	Maliyet
Herkesi onayla	%0	1,07
Eşik 0,5	%14	0,92
Eşik 0,09	**%81**	**0,50**

En düşük maliyetli karar, doğruluğu en yüksek karar değildir.

Eşik 0,09'da doğruluk %68'e düşüyor, ama maliyet neredeyse yarıya iniyor.

Eşik için basit bir kural

Olasılıklar güvenilirse, en iyi eşik maliyetlerden doğrudan hesaplanır

Temerrüt olasılığı p olan başvuran

Onaylarsak beklenen kayıp: $p \times 10$

Reddedersek beklenen kayıp: $(1 - p) \times 1$

Reddet: $p \times 10 > (1 - p) \times 1$

$$\text{eşik} = c_{YP} / (c_{YP} + c_{YN})$$

$$= 1 / (1 + 10) \approx 0{,}09$$

0,5 eşiği

İki hatanın maliyetinin eşit olduğu özel durumun cevabı.

Marjinal başvuran

Beklenen kaybın iki seçenekte eşitlendiği noktada ayrılır: tanıdık bir iktisat mantığı.

Koşul

Olasılıklar doğru olmalı; eşik test verisine değil eğitim verisine bakılarak seçilmeli.

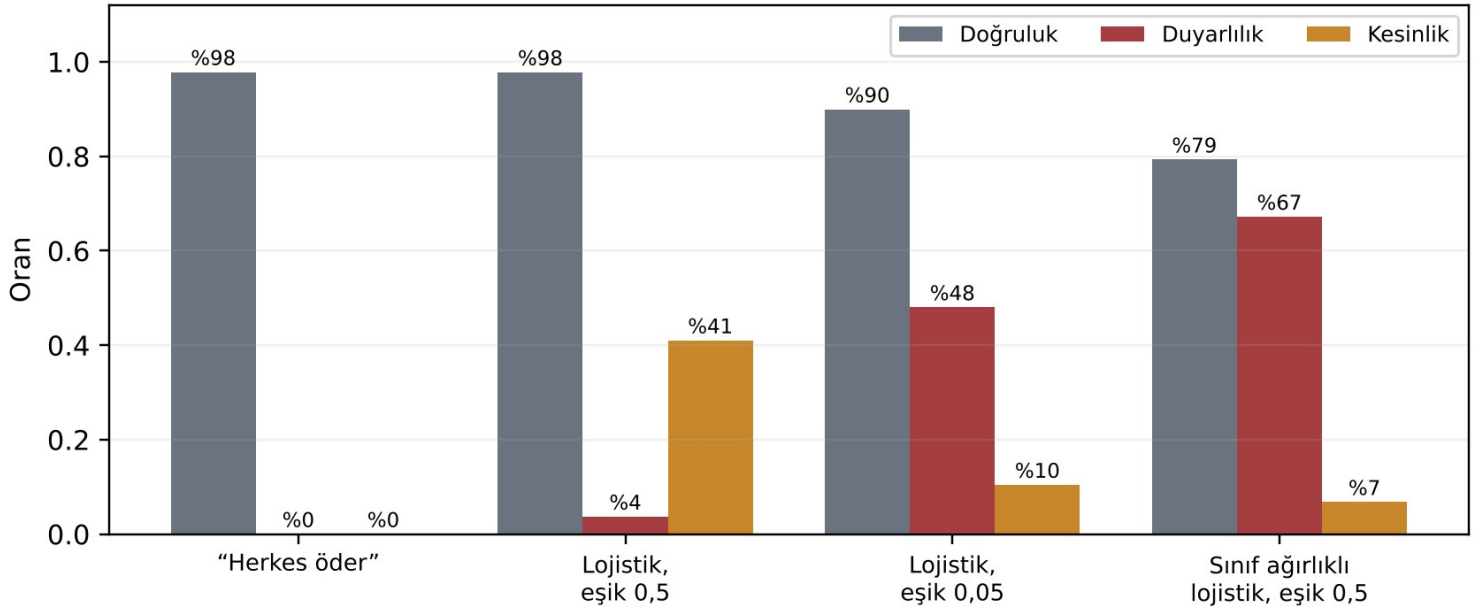
5

Dengesiz veri

Temerrüt, iflas, sahtecilik, kriz: ilgilendiğimiz olayların çoğu nadirdir

Temerrüt oranı %2 olunca

Eğitim verisinde 3.000 başvurudan yalnızca 63'ü temerrüde düşüyor



"Herkes öder"

%97,8 doğru; tek bir temerrüt bile yakalamıyor.

Model kötü değil

AUC 0,817. Sorun modelde değil, eşikte ve ölçüde.

Üç çare

Doğru ölçüye bak

Doğruluk yerine duyarlılık, kesinlik, AUC ya da beklenen maliyet. Bu bir ön koşul.

Eşiği ayarla

Model aynı, eşik 0,05.

Duyarlılık %4 → %48

Ağırlıklandır ya da yeniden örnekle

Azınlık sınıfındaki hatalara daha çok ağırlık ver; ya da çoğunluktan gözlem at, azınlığı çoğalt.

Duyarlılık %67

Ağırlıklandırmanın bedeli: olasılıklar bozulur

Gerçek temerrüt oranı %2,2

	Ortalama olasılık	AUC
Ağırlıksız model	%2,1	0,817
Sınıf ağırlıklı model	**%28,5**	0,815

AUC neredeyse aynı

Sıralama yeteneği değişmedi. Ağırlıklandırma olasılıkları yukarı kaydırarak eşiği dolaylı yoldan düşürdü.

Olasılık yanlış

Faiz fiyatlaması, karşılık ya da beklenen zarar için kullanılamaz. Önce eşik ayarını deneyin.

Dengeleme yalnızca eğitim verisine

Hafta 3'ün kuralı: önce böl, sonra hazırla



Test verisi dokunulmaz

Gerçek dünyadaki oranı korur. Dengelenmiş bir test verisinde ölçülen başarı, bankanın göreceği başarı değildir.



Katmanlı katlar

Çapraz doğrulamada her kat temerrüt oranını korumalı. 63 temerrütle bazı katlar neredeyse boş kalabilir.



Az gözlem az gözlemdir

63 temerrütten öğrenilebilecek bilgi sınırlı. Hiçbir dengeleme bunu yoktan var edemez.

Dengesiz veride doğruluk neredeyse hiçbir şey söylemez.

Önce sorun: hiç model kurmasaydım doğruluğum ne olurdu? Dengesizlik verinin hatası değil, dünyanın özelliğidir.

6

Uyarılar ve alıştırma

İktisatçı için

İktisatçı için üç uyarı



Eşik bir politika tercihi

Erken uyarı sisteminde kaçırılan kriz ile boşa çalan alarmın maliyetini kim belirliyor? Model bu tercihi yapamaz.



Zaman sırası

Geçmişle eğit, gelecekte test et. Ekonomik koşullar değişince temerrüt oranı da, doğru eşik de değişir.



Adillik

Model yasak bir özelliği ilişkili değişkenler üzerinden öğrenebilir. Hata oranlarını grup bazında raporlayın.

Alıştırma: hangi ölçüye bakmalı?

Gerekçesiyle cevap verin

1. “Temerrüt modelim %97 doğru.” Veride temerrüt oranı %3.
2. Sahtecilik tespiti: kaçırılan sahtecilik, durdurulan işlemden çok daha pahalı
3. Pazarlama ekibi yalnızca 1.000 kişiyi arayabiliyor
4. İki modeli karşılaştırıyoruz; eşiği sonra risk birimi seçecek
5. Sınıf ağırlıklı modelin olasılığı doğrudan faizi belirleyecek
6. Çeyreklik veriyle resesyon tahmini; 30 yılda birkaç resesyon çeyreği

Cevaplar

Ölçüyü, hangi hatanın daha pahalı olduğu belirler

1	“Temerrüt modelim %97 doğru.” Veride temerrüt oranı %3.	Hiçbir şey söylemiyor
2	Sahtecilik tespiti: kaçırılan sahtecilik, durdurulan işlemden çok daha pahalı	Duyarlılık, düşük eşik
3	Pazarlama ekibi yalnızca 1.000 kişiyi arayabiliyor	Kesinlik
4	İki modeli karşılaştırıyoruz; eşiği sonra risk birimi seçecek	AUC
5	Sınıf ağırlıklı modelin olasılığı doğrudan faizi belirleyecek	Sorunlu
6	Çeyreklik veriyle resesyon tahmini; 30 yılda birkaç resesyon çeyreği	Dengesiz + zaman + az gözlem

Bugünün beş mesajı

1. Lojistik regresyon doğrusal bir skoru S eğrisinden geçirip 0–1 arasında olasılık üretir: ekonometrideki logit.
2. Olasılık eşikle karara çevrilir. 0,5 eşiği, iki hatanın maliyeti eşitse doğrudur.
3. Karışıklık matrisi: doğruluk, duyarlılık, kesinlik, özgüllük farklı sorulara cevap verir ve aralarında ödünleşim vardır.
4. ROC tüm eşikleri gösterir; AUC eşikten bağımsız sıralama başarısıdır.
5. Dengesiz veride doğruluk yanıltır: önce doğru ölçü, sonra eşik; dengeleme olasılığı bozar ve yalnızca eğitimde yapılır.

Çıkış bileti

Üç dakika



1. Temerrüt oranı %5 olan bir veride “herkes öder” diyen kuralın doğruluğu ve duyarlılığı kaçtır?

2. Yanlış negatifin maliyeti yanlış pozitifin 4 katıysa eşik kaç olmalıdır?

3. Hâlâ kafamı karıştıran şey:

Gelecek hafta

Konu: Ağaçlar ve topluluk yöntemleri

- Veriyi sorularla bölen modeller
- Doğrusal olmayan ilişkiler ve etkileşimler
- Bugünkü ölçüler (karışıklık matrisi, ROC-AUC) aynen kullanılacak

Okuma

James vd. (2023)
An Introduction to Statistical Learning with Applications in Python

Bölüm 4.1–4.3: Classification, Logistic Regression
(bu haftanın ders kitabı karşılığı)

Kaynaklar

Elkan, C. (2001). The foundations of cost-sensitive learning. IJCAI, 973–978.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8).

Fuster, A. vd. (2022). Predictably unequal? The effects of machine learning on credit markets. Journal of Finance, 77(1).

Hanley, J. A. ve McNeil, B. J. (1982). The meaning and use of the area under a ROC curve. Radiology, 143(1).

James, G. vd. (2023). An Introduction to Statistical Learning with Applications in Python. Springer.

King, G. ve Zeng, L. (2001). Logistic regression in rare events data. Political Analysis, 9(2).

van den Goorbergh, R. vd. (2022). The harm of class imbalance corrections for risk prediction models. JAMIA, 29(9).