

Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead

Bora Mert Şahin - 181701004
Ecem Deniz Babaoğlu - 181201071

Giriş

- Şeffaflık
- Sorumluluk
- Örnek: şartlı tahliye, kefaletle serbest bırakılma, hava kirliliği modeli
- Cezai yargı, tıp, enerji güvenilirliği, finans ve diğer alanlar

- Explainable (Açıklanabilir) ML
- Interpretable (Yorumlanabilir) ML
- Alana özgü
- Alanla ilgili yapısal bilgilere uyar

Explainable(Açıklanabilir) ML ile ilgili temel sorunlar

Kapalı Kutu Modeller

- Bir insanın anaması için fazla karmaşık fonksiyonlar
- Şahsi, özel modeller

Explainable(Açıklanabilir) ML ile ilgili temel sorunlar

i- Modelin yorumlanabilir olması için doğruluğundan ödün verilmesi gereğinin düşünülmesi

- Düzgün yapılandırılmış, anlamlı özellikleri olan bir veri
- Computer vision
- Yorumlanamayan modeller, önemli kararlarda bilgi toplamak için faydalı olabilir

Explainable(Açıklanabilir) ML ile ilgili temel sorunlar

ii- Açıklanabilir makine öğrenimi yöntemleri, orijinal modelin hesapladıklarına sadık olmayan açıklamalar sağlar.

- Eğer açıklama modele sadık olsaydı modele gerek olmazdı.
- Örnek: suç tekrarı tahmini
 - Yaş + sabıka geçmişi
 - Irk
 - Doğru bir açıklama ama modele sadık değil
- “Açıklamalar” yerine “tahmin özetleri”, “özet istatistikler”, “eğilimler”

Explainable(Açıklanabilir) ML ile ilgili temel sorunlar

iii- Açıklamalar genellikle anlamsızdır veya kara kutunun ne yaptığını anlamak için yeterli ayrıntı sağlamaaz.

- Her iki model de doğru olsa bile,
- Açıklamanın modeldeki birçok bilgiyi çıkararak dahil etmemesi

Explainable(Açıklanabilir) ML ile ilgili temel sorunlar

iv- Kara kutu modelleri genellikle veritabanı dışındaki bilgilerin bir risk değerlendirmesiyle birleştirilmesi gereken durumlarla uyumlu değildir.

- Örnek: COMPAS modeli
- Güncel suçun ciddiyetinin dahil edilmemesi

Explainable(Açıklanabilir) ML ile ilgili temel sorunlar

v- Açıklamalı kara kutu modelleri, insan hatasına yatkın aşırı karmaşık bir karar yoluna sebep olabilir.

- Yazım hataları
- Sorumlu tutulmama

Interpretable(Yorumlanabilir) ML ile ilgili temel sorunlar

i- Şirketler, bir kara kutuya verilen fikri mülkiyetten kar elde edebilir.

IF	age between 18-20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21-23 and 2-3 prior offenses	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE		predict no arrest.

Figür 1: CORELS modeli

COMPAS	CORELS
black box 130+ factors might include socio-economic info expensive (software license), within software used in U.S. Justice System	full model is in Figure 3 only age, priors, (optional) gender no other information free, transparent

Figür 2: COMPAS ve CORELS karşılaştırması

Interpretable(Yorumlanabilir) ML ile ilgili temel sorunlar

ii- Yorumlanabilir modeller, hem hesaplama hem de alan uzmanlığı açısından inşa etmek için önemli çaba gerektirebilir.

Interpretable(Yorumlanabilir) ML ile ilgili temel sorunlar

iii- Kara kutu modelleri "gizli kalıpları" ortaya çıkarıyor gibi görünüyor

- Eğer bu gizli kalıplar önemli olsa yorumlanabilir modeller de bu kalıpları ortaya çıkarabilir

Sorumlu Makine Öğrenimi Yönetimini Teşvik Etmek

Avrupa Birliği Genel Veri Koruma Yönetmeliği

- “Açıklama hakkı”
- “Veri öznesi, kendisini ilgilendiren hukuki sonuçlar doğuran veya benzer şekilde önemli ölçüde etkileyen profil oluşturma da dahil olmak üzere, yalnızca otomatik işlemeye dayalı bir karara tabi tutulmama hakkına sahip olacaktır.”



Sorumlu Makine Öğrenimi Yönetişimini Teşvik Etmek

Getirilebilecek Kurallar

- Yüksek riskli durumlarda, eğer aynı performansa sahip yorumlanabilir bir model bulunuyorsa, kapalı kutu modeller kullanılmamalıdır.
- Kapalı kutu model hazırlayan organizasyonların yorumlanabilir makine öğrenmesi modellerinin doğruluk metriklerini de paylaşması zorunlu tutulmalıdır.

Yorumlanabilir Makine Öğrenmesinin Algoritmik Zorlukları

Zorluk 1: Optimal Mantıksal Modeller Oluşturmak

- ‘or’, ‘and’, ‘if-then’
- Heuristic metodlar:
 - Mantıksal şartları tek tek eklemek
 - Modelin sonunda gereksiz şartları seçip kırpmak
- El ile yapılmış gibi gözüken yüksek doğruluklu makine öğrenmesi modelleri

Table 1 | Machine learning model from the CORELS algorithm

IF	age between 18-20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21-23 and 2-3 prior offences	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest	

This model from ref.³⁹ is the minimizer of a special case of equation (1) discussed later in the challenges section. CORELS' code is open source and publicly available at <http://corels.eecs.harvard.edu/>, along with the data from Florida needed to produce this model.

Yorumlanabilir Makine Öğrenmesinin Algoritmik Zorlukları

Zorluk 1: Optimal Mantıksal Modeller Oluşturmak

- CORELS
 - Kural listesi uzayını küçültten teorem setleri
 - Uzay içinde aramayı kolaylaştıran yöntemler
 - İşlemleri ve simetrileri kontrol altında tutan özel veri yapıları

Table 1 | Machine learning model from the CORELS algorithm

IF	age between 18-20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21-23 and 2-3 prior offences	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest	

This model from ref.³⁹ is the minimizer of a special case of equation (1) discussed later in the challenges section. CORELS' code is open source and publicly available at <http://corels.eecs.harvard.edu/>, along with the data from Florida needed to produce this model.

Yorumlanabilir Makine Öğrenmesinin Algoritmik Zorlukları

Zorluk 2: Optimum Seyrek Puanlama Skorlama Oluşturmak

- **Burgess kriminoloji modeli (1928)**
 - İlk skorlama sistemi

Table 3 | Scoring system for risk of recidivism

1.	Prior arrests ≥ 2	1 point	...
2.	Prior arrests ≥ 5	1 point	+...
3.	Prior arrests for local ordinance	1 point	+...
4.	Age at release between 18 to 24	1 point	+...
5.	Age at release ≥ 40	-1 point	+...
		Score	= ...
Score	-1 0 1 2 3 4		
Risk (%)	11.9 26.9 50.0 73.1 88.1 95.3		

This system is from ref. ²¹, which was developed from refs. ^{29,46}. The model was not created by a human; the selection of numbers and features come from the RiskSLIM machine learning algorithm.

Yorumlanabilir Makine Öğrenmesinin Algoritmik Zorlukları

Zorluk 3: Spesifik İlgi Alanları İçin Yorumlanabilirlik Tanımlama ve Yöntemler Oluşturma

- Computer Vision
 - ‘This look like that’

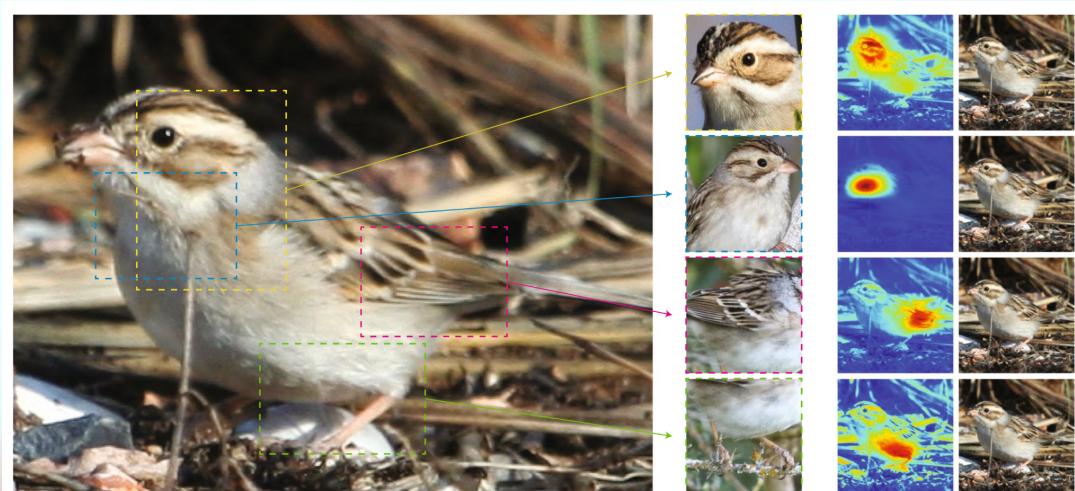


Fig. 3 | Image from the authors of ref. ⁴⁸, indicating that parts of the test image on the left are similar to prototypical parts of training examples. The test image to be classified is on the left, the most similar prototypes are in the middle column, and the heatmaps that show which part of the test image is similar to the prototype are on the right. We included copies of the test image on the right so that it is easier to see to what part of the bird the heatmaps are referring. The similarities of the prototypes to the test image are what determine the predicted class label of the image. Here, the image is predicted to be a clay-coloured sparrow. The top prototype seems to be comparing the bird's head to a prototypical head of a clay-coloured sparrow, the second prototype considers the throat of the bird, the third looks at feathers, and the last seems to consider the abdomen and leg. Credit: Image constructed by Alina Barnett, Duke University

Yorumlanabilir Makine Öğrenmesinin Algoritmik Zorlukları

Zorluk 3: Spesifik İlgi Alanları İçin Yorumlanabilirlik Tanımlama ve Yöntemler Oluşturma

- 2018 FICO Explainable ML Challenge
 - Accuracy/interpretability trade-off
 - Kredi skorlama
- Rashomon setleri:
 - Birbirlerine benzer performans gösteren neredeyse eşit modellerin bulunduğu setler

Dinlediğiniz için Teşekkürler

Bora Mert Şahin - 181701004
Ecem Deniz Babaoğlu - 181201071