Sinopfbd, 2024; 9(1): 154-166 **https://doi.org/10.33484/sinopfbd.1434260**

Sinop Üniversitesi Fen Bilimleri Dergisi
Sinop Uni J Nat Sci
E-ISSN: 2564-7873

https://dergipark.org.tr/tr/pub/sinopfbd

Çoklu Doğrusal Bağlantılı Nadir Olayların Modellenmesinde Lasso ve Ridge Regresyon ile Boosting Algoritmalarının Performans Karşılaştırması

Olcay ALPAY

How to cite: Alpay, O. (2024). Çoklu doğrusal bağlantılı nadir olayların modellenmesinde Lasso ve Ridge regresyon ile Boosting algoritmalarının performans karşılaştırması. *Sinop Üniversitesi Fen Bilimleri Dergisi, 9*(1), 154-166. https://doi.org/10.33484/sinopfbd.1434260

Araştırma Makalesi

Sorumlu Yazar
Olcay ALPAY
olcayb@sinop.edu.tr

Yazara ait ORCID
O.A: 0000-0003-1446-0801

Received: 09.02.2024
Accepted: 21.05.2024

Öz

Bu çalışma, iki durumlu olayları modellemek için kullanılan makine öğrenmesi tekniklerinde karşılaşılan nadirlik ve "çoklu doğrusal bağlantı" ya da sadece "çoklu bağlantı" olarak tanımlanan sorunu ele alınmaktadır. Çoklu doğrusal bağlantı (ÇDB), bağımsız değişkenler arasında bir ya da birden fazla kuvvetli doğrusal bağımlılık olma durumudur ve bir sorun olarak ortaya çıkar. Üzerinde çalışılan veri içerisinde çoklu doğrusal bağlantı probleminin var olması regresyon katsayılarının varyanslarının büyümesi gibi olumsuz bir sonuca sebebiyet verir. Bu çalışmada, Lasso ve Ridge Regresyon ile GradientBoost, XGBoost, LightGBM ve AdaBoost gibi artırma algoritmaları içeren düzenleme ve ölçeklendirme tekniklerinin, çoklu doğrusal bağlantılı nadir olayların modellenmesinde, algoritmaların performanslarını karşılaştırmak için detaylı bir simülasyon çalışması sunulmaktadır. Simülasyon çalışmasında, verideki dengesizliği ortadan kaldırmak amacıyla yeniden örnekleme yöntemleri kullanılarak sonuçlara etkisi Hata Kareler Ortalaması (*HKO*), R^2, Hassasiyet (Precision-Prec), Duyarlılık (Recall-Rec) ve Eğri Altında Kalan Alan (Area Under the Curve-AUC) gibi performans metrikleri ve İşlem Karakteristik Eğrisi (Receiver Operating Characteristic- ROC) grafikleri ile araştırılmaktadır. Sonuçlar Lasso, Ridge ve Boosting algoritmalarının ÇDB'ya sahip nadir olayların modellenmesinde hangi yöntemin uygun olduğunu belirlemek açısından katkı sunmaktadır.

Anahtar Kelimeler: Lasso regresyon, Ridge regresyon, Boosting algoritmaları, performans metrikleri, yeniden örnekleme teknikleri

Performance Comparison of Lasso and Ridge Regression and Boosting Algorithms for Modeling Rare Events with Multicollinearity

Sinop Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü, Sinop, Türkiye

Abstract

This study examines the issues of rarity and multicollinearity in machine learning techniques used to model binary events. Multicollinearity (MC) is the presence of strong linear dependencies among independent variables, which poses a problem. In the context of the data being studied, the existence of multicollinearity leads to undesired consequences such as an enlargement of the variances of the regression coefficients. This study presents a simulation comparing the performance of algorithms in modelling multicollinear and rare events. Regularization and scaling techniques such as Lasso and Ridge Regression, as well as Boosting algorithms like GradientBoost, XGBoost, LightGBM, and AdaBoost are

utilized. The impact of resampling methods to reduce data imbalance is also investigated using performance metrics such as Mean Squared Error (MSE), R^2, Precision (Prec), Recall (Rec) and AUC values, along with ROC curves. The results help to determine the appropriate method for modelling rare events with multicollinearity and provide insight into the performance of Lasso, Ridge and Boosting algorithms.

Keywords: Lasso regression, Ridge regression, Boosting algorithms, performance metrics, resampling techniques

Giriş

Çoklu doğrusal bağlantı (ÇDB), çoklu regresyon modellinin temel varsayımlardan birinin ihlalidir ve regresyon modellerindeki bağımsız değişkenler arasında yüksek bir ilişki olduğunda ortaya çıkan bir sorundur. Veri içerisinde çoklu doğrusal bağlantı problemi olması, regresyon katsayılarının varyanslarının büyümesine sebep olur, dolaysıyla tahmin edilen parametrelerin güvenilirliği azalır [1]. İki durumlu olayların modellenmesinde kullanılan birçok farklı makine öğrenmesi algoritması bulunmaktadır. Bunlardan en yaygın olanı ise lojistik regresyondur. Özellikle, nadir olayların modellenmesinde lojistik regresyon, parametrelerinin değerlerini olması gerekenden daha düşük tahmin etme eğilimindedir [2]. Nadir görülen olaylara örnek olarak savaşlar, çok şiddetli depremler, büyük çaplı salgınlar verilebilir [3]. Lojistik regresyonda da çoklu doğrusal bağlantı sorunlarıyla karşılaşılabilir. Dolayısıyla, çoklu doğrusal bağlantı, lojistik regresyon modelinde bağımsız değişkenler arasındaki ilişkiyi etkileyebilir ve nadirlik yüzünden olması gerekenden daha düşük değerli tahmin edilen model parametrelerinin güvenilirliğini daha da azaltabilir. Bu nedenle, lojistik regresyon modelleri oluştururken çoklu doğrusal bağlantı sorununa dikkat etmek ve etkilerini azaltmak önemlidir. Ridge veya Lasso regresyon gibi düzenleme teknikleri, modeldeki katsayıları sıfıra yaklaştırarak veya sıfıra eşitleyerek çoklu doğrusal bağlantı sorununu azaltabilir. Bu şekilde, lojistik regresyon modelinin performansı ve güvenilirliği artırılabilir. Shrivastava ve ark. [4] finansal sistemde önemli bir role sahip bankalar üzerine bir çalışma yürütmüşlerdir. Hindistan'da batan bankalara ait veri setindeki dengesizliği dikkate alarak ve SMOTE yeniden örnekleme tekniğini kullanarak Lasso regresyon, AdaBoost ve başka makine öğrenmesi algoritmaları ile bankaların başarısızlığını tahmin etmek için uygun makine öğrenme tekniğinin seçilmesi üzerine bir yaklaşım sunmayı amaçlamışlardır. 2020 yılında Rochayani ve ark. [5] yüksek boyutlu ve dengesiz sınıflı gen veri setlerinde kanser sınıflandırması için az örnekleme (undersampling) ve gen seçimi üzerine bir çalışma gerçekleştirmişlerdir. Veri setini dengelemek için rastgele az örnekleme yöntemini kullanmışlar ve gen seçimi için Lasso regresyonu tercih etmişlerdir. Ridge regresyon ile ilgili yapılan literatür taramasında, Ridge regresyonun dengeli veri setlerindeki uygulamalarına sıkça rastlanmış, ancak nadir olaylar üzerinde Sıradan Ridge Regresyonu (Ordinary Ridge Regression) kullanan bir çalışma ile karşılaşılmamıştır. Değişkenler arasında çoklu doğrusal bağlantı olduğunda kullanılabilecek diğer bir yaklaşım, makine öğrenmesi çerçevesinde incelenen Boosting algoritmalarıdır. Bu konu ile ilgili literatürde yapılan çalışmalardan bazıları şöyledir: Cahyana ve ark. [6] yaptıkları çalışmada göğüs kanseri verisinde başta SMOTE olmak üzere çeşitli yeniden

örnekleme yöntemlerini GradientBoost algoritmasında kullanmış ve performanslarını değerlendirmişlerdir. Zaten yüksek başarı gösteren GradientBoost algoritmasının performansının aşırı örnekleme (oversampling) ile daha da arttığı yaptıkları çalışmada gösterilmiştir. Tanha ve ark. [7] yaptıkları çalışmada ise içlerinde GradientBoost, XGBoost, LightGBM ve AdaBoost'un da olduğu Boosting algoritmalarının dengesiz veri kümeleri üzerindeki performansını analiz etmek için kapsamlı bir deneysel karşılaştırma yapmışlardır. Ashraf ve ark. [8] yaptıkları çalışmada, XGBoost algoritmasında SMOTE, Tomek bağlantısı ve başka yeniden örnekleme yöntemlerini kullanarak yanlış yönden araç kullanımına bağlı kazalar için yüksek riskli yol kesimlerinin belirlenmesi üzerine bir çalışma yürütmüşlerdir. Prec, Rec, AUC ve ROC sonuçlarını kullanarak performans değerlendirmesini sunmuşlardır. Bu çalışma ile literatürde daha önce nadir olaylarda kullanılmamış Ridge regresyonu ilk kez incelemek amaçlanmıştır. Aynı zamanda, Boosting algoritmalarının Lasso ve Ridge regresyon ile farklı yeniden örnekleme teknikleri ve örneklem büyüklüklerinde performanslarını karşılaştırılarak literatüre katkı sağlamak hedeflenmiştir. Çalışmanın takip eden bölümünde nadir olayları modellemek için kullanılacak Lasso ve Ridge Regresyon ile Boosting algoritmaları, SMOTE, Tomek Bağlantısı ve SMOTETomek yeniden örnekleme teknikleri, son olarak da sınıflandırma performansını değerlendirmek için kullanılan ölçümler açıklanacaktır. Bir sonraki bölümde simülasyon çalışması verilecek ve son bölümde ise elde edilen sonuçlar değerlendirilecektir.

Metodoloji

Lasso Regresyon

Lasso regresyon, katsayıların mutlak değerlerinin toplamının bir sabitten küçük olması koşuluyla artık kareler toplamını en aza indirir. Bu yöntem, bazı regresyon katsayılarının büyüklüğünü tam sıfır olarak üreterek sınırlar ve modelin karmaşıklığını azaltır.

$$\hat{\beta}_{lasso} = argmin \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_j \beta_j x_{ij} \right)^2 + \lambda \sum_j |\beta_j| \right\}$$

Buradaki $\lambda \geq 0$ büzülme miktarını kontrol eden ayarlama parametresidir [9]. $\sum_j |\beta_j|$'nin kullanılması parametre tahminlerini sıfırlamaya zorladığı için modelin değişken sayısını azaltıp, daha az değişkene sahip bir model seçmeyi ve tahmin yapmayı sağlar [10].

Ridge Regresyon

Hoerl ve Kennard tarafından 1970 yılında tanıtılan Ridge tahmin edicisi, açıklayıcı değişkenlerin arasındaki ÇDB sorununu çözmek ve böylece tahmin edicilerin varyanslarını azaltmak için önerilmiştir [11]. Bu yöntem büyük katsayıları küçültür, ancak sıfıra indirmez.

$$\hat{\beta}_{ridge} = argmin\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_j \beta_j x_{ij}\right)^2 + \lambda \sum_j \beta_j^2\right\}$$

Burada λ ayarlama parametresidir.

Boosting Algoritmaları

"Boosting", yanlılığı ve varyansı azaltmak amacıyla zayıf öğrenicileri güçlü öğrenicilere dönüştüren bir topluluk öğrenme tekniğidir [12]. Çalışmada kullanılan bu ailenin en yaygın tekniklerinden aşağıda bahsedilmiştir.

GradientBoost (Gradient Boosting)

GradientBoost, kendinden önceki zayıf modellerdeki (zayıf öğrenenler) yanlışlıkları hesaba katarak modeli kapsamlı bir şekilde ayarlayan ve daha güçlü tahmin modeli oluşturan bir makine öğrenme algoritmasıdır [13]. Algoritmanın en önemli dezavantajı, önceki ağaçların hatasını azaltabilmek amacıyla sürekli yeni ağaçlar oluşturması ve küçük veri setlerinin bile eğitiminin çok zaman almasıdır [14].

GradientBoost Algoritması [15]:
Adım 1: Başlangıç tahmini yap
Adım 2: Kayıp fonksiyonunu kullanarak negatif gradiyenti hesapla
Adım 3: Hatayı düzeltmek için yeni öğrenici oluştur
Adım 4: Öğrenme oranı ile ağırlıklandır ve hatayı düzelt
Adım 5: Sonuç tahmini yap

XGBoost (eXtreme Gradient Boosting)

XGBoost, Chen tarafından 2016 yılında tanıtılan GradientBoost'un optimize edilmiş bir versiyonudur. XGBoost, geleneksel GradientBoost algoritmasının eğitim aşamasındaki aşırı uyumunu, ayrık ve eksik değerlerini kontrol etmek için verimli bazı özelliklerin eklenmiş versiyonudur. XGBoost, uygulama düzeyinde mükemmel iyileştirmeler yapmış, çok büyük veri kümeleri için uygulanabilir bir algoritmadır [7]. XGBoost'un GradientBoost'a göre en büyük avantajı, ağaçların ayrı ayrı birden fazla çekirdek kullanılarak oluşturulmasıdır. Bu sayede veriler, arama süreleri kısalacak şekilde düzenlenir [14]. Ayrıca, güçlü genelleme yeteneği ve yüksek genişletilebilirlik avantajlarına sahiptir [16].

LightGBM (Light Gradient Boosting Machine)

LightGBM, Ke ve ark. tarafından 2017'de yüksek boyutlu özellik uzayı veya büyük boyutlu verilerde GradientBoost'un verimlilik ve ölçeklenebilirlik sorununu çözmek için önerilmiştir [7]. Algoritma,

hesaplama gücünü ve tahmin doğruluğunu iyileştirmek için temel olarak histogram algoritmasını ve diğer algoritmaları kullanır [17].

LightGBM Algoritması [18]:

Adım 1: Veri hazırlığı yap
Adım 2: Modeli başlat
Adım 3: Özelliklerin boyutunu azalt
Adım 4: Örneklerin gradiyentlerini hesapla
Adım 5: Büyük gradiyentli örnekleri tut
Adım 6: Histogram oluştur
Adım 7: Optimal segmentasyonu bulmak için histogramı çaprazla
Adım 8: Tahmin yap

AdaBoost (Adaptive Boosting)

AdaBoost algoritması Boosting algoritmaları içinde en iyi olanlardan biridir [19]. AdaBoost'ta ilk olarak, eğitim veri kümesine eşit ağırlık atanır. Daha sonra, AdaBoost minimum ağırlığa sahip en iyi özelliği dikkate alan çok kısa bir ağaç oluşturarak başlatılır. Ağırlıklar tüm özellikler için aynı olduğundan, ilk ağaç ilk özelliği dikkate alır, daha sonra tüm ağaçlar diğer özelliklere göre oluşturulur. Tüm ağaçlar için Gini endeksi hesaplanır ve tüm düğümlerin tüm ağaçlar üzerindeki ağırlığı değerlendirilir. Ardından, yeni ağırlık eski ağırlığın yerini alır. Sonraki aşamada, AdaBoost yapılan yanlışlara odaklanarak yeni ağaçları bir önceki ağacın yaptığı hataya göre oluşturur. Yani, bu ağaç önceki ağaçlardan biraz daha iyi ve biraz daha büyüktür. AdaBoost bu prosedürde uygun olana kadar sürekli olarak yeni ağaçlar oluşturur [6].

AdaBoost Algoritması [15]:

Adım 1: Başlangıç ağırlıklarını ata
Adım 2: Veri setindeki örneklerin ağırlıklarına göre bir öğrenici oluştur
Adım 3: Öğrenciyi eğit
Adım 4: Hatayı hesapla
Adım 5: Ağırlıkları belirle
Adım 6: Sonuç tahminini yap

Yeniden Örnekleme Teknikleri

Yeniden örnekleme, sınıf dengesizliği sorununu çözmek için yaygın olarak kullanılan bir yöntemdir. Örneklemenin amacı, geleneksel sınıflandırıcıların daha dengeli bir sınıf dağılımına sahip bir veri seti oluşturarak çoğunluk ve nadir sınıflar arasındaki karar sınırını daha doğru bir şekilde yakalamasına olanak sağlamaktır [20].

SMOTE

SMOTE (sentetik azınlık aşırı örnekleme tekniği), nadir sınıfı aşırı örnekleyerek verileri dengeleyen önemli bir yaklaşımdır [21]. Aşırı örnekleme teknikleri, veri kümesindeki nadir olayların temsilini iyileştirebilse de aşırı uyumla ilgili sorunlara yol açabilir [20].

Tomek Bağlantısı

Dengeli veri seti elde etmek amacıyla çoğunluk sınıfının bazı gözlemlerini hariç tutarak veri setini rastgele örneklemeyi ifade eder. Bu şekilde nadir olaylar veri setinde daha iyi temsil edilir [22]. Ancak veri setinin nadirlik seviyesine bağlı olarak çok sayıda verinin silinmesine sebep olabilir.

SMOTETomek

SMOTETomek, SMOTE ile Tomek bağlantısını birleştiren hibrit bir yaklaşımdır. SMOTE, nadir olayların sayısını baskın sınıfın sayısına eşit olana kadar çoğaltan ve Tomek bağlantısı ile araştırmada veri işleme sonrasında temizleme adımını gerçekleştirilen bir alt örnekleme yöntemidir [23]. Bu hibrit yaklaşım, ana sınıfın azaltılmasının önemli bir bilgi kaybına yol açmamasını sağlarken aynı zamanda aşırı örneklemeden kaynaklanabilecek aşırı uyum sorununu da önler [20]. Genel olarak yeniden örnekleme yöntemlerinin çalışma prensipleri özetlenecek olursa, SMOTE tekniğinde sınıfların dengelenmesi, nadir sınıf için üretilen sentetik verilerle çoğunluk sınıfının yoğunluğuna ulaşarak sağlanırken, Tomek bağlantısında çoğunluk sınıfın veri sayısının azaltılarak nadir sınıfa yaklaştırılmasıyla gerçekleştirilir. Hibrit yaklaşımda ise SMOTE ile veri çoğaltma ve Tomek bağlantısı ile veri azaltma durumu söz konusudur [24].

Performans Metrikleri

Sınıflandırıcılar aracılığıyla ikili sınıflandırma yapılırken, doğru sınıflandırma oranının ölçülerek performans değerlendirilmesi yapılmasına ihtiyaç vardır. Bu amaçla nadir olan sınıfın tahmin performansını belirlemek için HKO, $R^2, Prec, Rec$ ve AUC gibi performans metrikleri bu çalışmada kullanılmıştır.

$\hat{y}_i$, i. örneklemin tahmini, y_i karşılık gelen gerçek değeri ve n test veri setinin boyutunu göstermek üzere HKO ve R^2 aşağıdaki gibi hesaplanmaktadır.

$$HKO(y, \hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2$$

ve

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Diğer metrikleri hesaplamak için kullanılan karışıklık matrisi ise şu şekilde tanımlanır:

		Tahmin edilen	
		1	0
Gözlenen	1(nadir sınıf)	DP	YN
	0	YP	DN

Prec ve Rec metrikleri aşağıdaki gibidir.

$$Prec = \frac{DP}{DP + YP}$$

$$Rec = \frac{DP}{DP + YN}$$

DP: Doğru Pozitif, **YP**: Yanlış Pozitif, **DN**: Doğru Negatif ve **YN**: Yanlış Negatifi ifade etmektedir. ROC eğrisi, duyarlılık ile 1-özgüllük arasındaki ilişkiyi gösteren bir haritadır [25]. Çeşitli kesme değerlerine göre elde edilen yanlış pozitif oranlara karşı hassasiyetlerin grafiğini çizerek bir sınıflandırıcının tanısal yeteneğini değerlendirmek için kullanılır [26]. ROC güçlü bir performans ölçütü olmasına rağmen eğri altındaki alanın (AUC) nadir olay sınıflandırması için daha uygun olduğu öne sürülmüştür [27]. AUC, tüm ROC eğrilerinin iyi bir özetini sunar [28].

Simülasyon Çalışması

İkili olayların modellenmesi için kullanılacak olan algoritmaların deneysel performanslarının değerlendirilmesinde, kesişme katsayısı olan β_0, ilgilenilen olayın veri seti içindeki oranını belirlediğinden bu değeri en iyi şekilde temsil edebilmek adına örneklem ölçümü en az $n = 1000$ ve en çok $n = 5000$ olarak seçilmiştir. İlgilenilen nadir olay "1" ile gösterilmekte olup örneklem içindeki oranı %15'tir. Nadir olayın oluşması ve %15'lik nadirlik seviyesine ulaşılabilmesi için, 1000 birimlik örneklem büyüklüğünde $\beta_0 = -21.4$ ve 5000 birimlik örneklem büyüklüğünde ise $\beta_0 = -26.1$ olarak belirlenmiştir. Standartlaştırılmış $\beta_i = 1$ ve $X_i \sim N(0,1)\ i = 1,2,\dots,10$ normal dağılıma sahip 10 açıklayıcı değişkenden oluşan simülasyon tasarımında, değişkenler arasında ilişki oluşturabilmek amacıyla bazı değişkenler arasında doğrusal ilişki kurulmuş ve Tablo 1'de verilen ilişki değerleri elde edilmiştir. Daha sonra yukarıdaki ayarlara sahip ikili olay şu şekilde oluşturulmuştur:

$$y(\pi(x)) = \begin{cases} 0, & eğer\ \pi < u \\ 1, & d.d. \end{cases}$$

burada *u* yapay değişkendir.

Python Spyder 5.4.3'te yazılan ve 100 tekrar ile gerçekleştirilen simülasyonda veri setleri rasgele olarak eğitim ve test setlerine 2/3:1/3 oranında ayrılmıştır. İlk olarak dengesiz dağılıma sahip, sonrasında ise imblearn [29] kütüphanesinden yararlanarak yeniden örnekleme teknikleriyle dengeli hale getirilmiş

verilerin, eğitim setinde model öğrenimi gerçekleştirildikten sonra test setinden tahmin değerleri elde edilmiş ve gerekli metrikler hesaplanmıştır. Simülasyon sayısı kadar tekrarlanan çalışma sonunda elde edilen ortalama değerler Tablo 2-4'te sunulmuştur.

Tablo 1 *İlişki matrisi*

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1	0.60	0.15	0.20	0.41	0.25	0.14	0.72	0.12	0.09
X_2	0.60	1	0.18	0.23	0.07	0.27	0.31	0.14	0.07	0.01
X_3	0.15	0.18	1	0.05	0.11	0.28	0.15	0.02	0.40	0.31
X_4	0.20	0.23	0.05	1	0.08	0.16	0.07	0.30	0.19	0.01
X_5	0.41	0.07	0.11	0.08	1	0.04	0.19	0.12	0.10	0
X_6	0.25	0.27	0.28	0.16	0.04	1	0	0.01	0.72	0.08
X_7	0.14	0.31	0.15	0.07	0.19	0	1	0.11	0.08	0.17
X_8	0.72	0.14	0.02	0.30	0.12	0.01	0.11	1	0	0
X_9	0.12	0.07	0.40	0.19	0.10	0.72	0.08	0	1	0
X_{10}	0.09	0.01	0.31	0.01	0	0	0.17	0	0	1

Tablo 2'de algoritmaların oluşturdukları modellere ait Hata Kareler Ortalaması (*HKO*) ve Tablo 3'te model açıklayıcılıkları (R^2) sunulmaktadır. Tablolar detaylı olarak incelendiğinde örneklem ölçümü büyüdüğünde, Lasso regresyonun *HKO* değerinin yükseldiği ve buna bağlı olarak model açıklayıcılığının bir miktar düştüğü görülmektedir. Ridge regresyon açısından değerlendirme yapıldığında ise örneklem ölçümünün artması *HKO* ve R^2 açısından çok büyük farklılıklara sebep olmamıştır. Boosting algoritmalarına bakılacak olursa yeniden örnekleme yapılmadan veya yeniden örnekleme tekniklerinin tümü için en düşük *HKO* ve en yüksek R^2 değerine sahip algoritmanın LightGBM olduğu görülmektedir. GradientBoost en iyi performansını veri dağılımındaki dengenin çoğunluk sınıfın sayısını nadir olay sayısına yaklaştırmayı hedefleyen Tomek bağlantısı kullanıldığında elde etmiştir. Bu durum XGBoost ve AdaBoost için de benzer şekildedir.

Tablo 2 *Algoritmaların örneklem ölçümlerine bağlı HKO değerleri*

HKO	**Yeniden Örnekleme Yok**		**SMOTE**		**Tomek Bağlantısı**		**SMOTETomek**	
Model	**n=1000**	**n=5000**	**n=1000**	**n=5000**	**n=1000**	**n=5000**	**n=1000**	**n=5000**
Lasso	0.088	0.089	0.108	0.110	0.086	0.088	0.107	0.111
Ridge	0.074	0.074	0.105	0.104	0.073	0.073	0.104	0.105
GradientBoost	0.030	0.027	0.035	0.038	0.028	0.027	0.034	0.037
XGBoost	0.033	0.028	0.036	0.033	0.032	0.028	0.033	0.032
LightGBM	**0.026**	**0.024**	**0.027**	**0.028**	**0.025**	**0.025**	**0.025**	**0.027**
AdaBoost	**0.026**	0.041	0.029	0.047	0.025	0.041	0.028	0.048

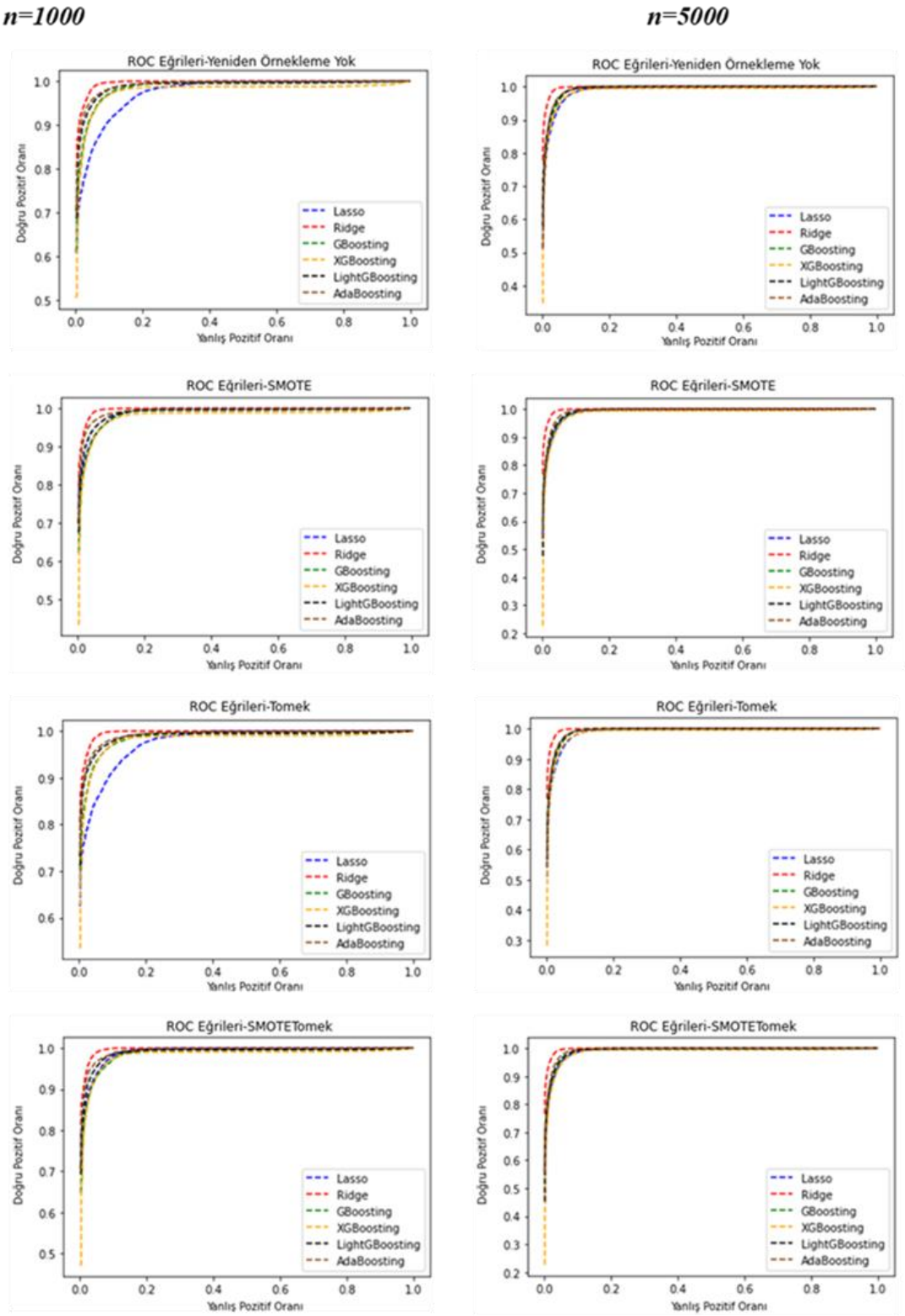
Tablo 3 *Algoritmaların örneklem ölçümlerine bağlı R^2 değerleri*

R^2	Yeniden Örnekleme Yok		SMOTE		Tomek Bağlantısı		SMOTETomek	
Model	**n=1000**	**n=5000**	**n=1000**	**n=5000**	**n=1000**	**n=5000**	**n=1000**	**n=5000**
Lasso	0.312	0.303	0.150	0.133	0.322	0.308	0.137	0.134
Ridge	0.418	0.424	0.175	0.180	0.424	0.421	0.166	0.181
GradientBoost	0.763	0.790	0.722	0.706	0.776	0.789	0.731	0.707
XGBoost	0.735	0.784	0.716	0.742	0.750	0.779	0.733	0.748
LightGBM	**0.796**	**0.809**	**0.785**	**0.784**	**0.806**	**0.806**	**0.797**	**0.787**
AdaBoost	0.795	0.678	0.775	0.628	0.805	0.674	0.775	0.625

Tablo 4'te ise algoritmaların doğru sınıflandırma performanslarını belirlemek amacıyla hesaplanan metrikler verilmiştir. Lasso regresyon metrikler açısından en yüksek skorlarını örneklem ölçümünden bağımsız olarak SMOTE yeniden örneklemesinde elde etmiştir. Ridge regresyon, Prec metriği açısından SMOTETomek, Rec açısından Tomek bağlantısı ve AUC açısından ise tüm örneklem ölçümleri ve tüm yeniden örnekleme tekniklerinde en iyi performansı göstermiştir. GradientBoost' un tüm performans metrikleri için örneklem ölçümü ve yeniden örnekleme tekniği fark etmeksizin birbirine yakın değerler verdiği görülmektedir. Bu durum XGBoost ve LightGBM içinde benzerdir. Prec ve Rec metriklerine göre AdaBoost yöntemi neredeyse tüm örneklem ölçümü ve örnekleme tekniklerinde iyi performans değerlerine sahiptir.

Tablo 4 *Algoritmaların örneklem ölçümlerine bağlı performans metrik değerleri*

		Yeniden Örnekleme Yok		SMOTE		Tomek Bağlantısı		SMOTETomek	
Model		**n=1000**	**n=5000**	**n=1000**	**n=5000**	**n=1000**	**n=5000**	**n=1000**	**n=5000**
Lasso	**Prec**	0.842	0.869	0.934	0.934	0.858	0.872	0.938	0.933
	Rec	0.855	0.851	0.888	0.882	0.856	0.853	0.895	0.881
	AUC	0.975	0.990	0.990	0.991	0.976	0.990	0.991	0.991
Ridge	**Prec**	0.925	0.926	0.940	0.935	0.929	0.931	0.942	0.935
	Rec	0.911	0.917	0.886	0.867	0.917	0.923	0.890	0.867
	AUC	**0.996**	**0.998**	**0.997**	**0.998**	**0.997**	**0.998**	**0.997**	**0.998**
GradientBoost	**Prec**	0.959	0.965	0.956	0.957	0.962	0.965	0.958	0.957
	Rec	0.960	0.966	0.951	0.946	0.962	0.965	0.954	0.947
	AUC	0.987	0.993	0.987	0.992	0.987	0.993	0.988	0.992
XGBoost	**Prec**	0.957	0.962	0.956	0.958	0.960	0.962	0.959	**0.969**
	Rec	0.957	0.962	0.953	0.955	0.959	0.961	0.956	0.956
	AUC	0.978	0.987	0.979	0.985	0.981	0.987	0.981	0.986
LightGBM	**Prec**	0.969	0.966	0.965	0.963	0.971	0.966	0.968	0.965
	Rec	0.969	0.966	0.964	0.961	0.971	0.966	0.967	**0.962**
	AUC	0.989	0.993	0.989	0.991	0.988	0.993	0.989	0.992
AdaBoost	**Prec**	**0.972**	**0.968**	**0.972**	**0.965**	**0.973**	**0.968**	**0.974**	0.965
	Rec	**0.972**	**0.967**	**0.971**	**0.959**	**0.973**	**0.967**	**0.973**	0.959
	AUC	0.992	0.993	0.993	0.993	0.992	0.993	0.994	0.994



Şekil 1 *Örneklem ölçümlerine bağlı ROC eğrileri*

Şekil 1'de verilen ROC eğrileri incelendiğinde Ridge regresyonun hem gerçek pozitifleri hem de gerçek negatifleri en doğru şekilde sınıflandırdığı görülmektedir. Bu durum Tablo 4'teki AUC sonuçlarını desteklemektedir.

Sonuçlar

Farklı regresyon ve Boosting algoritmalarının performansları incelendiğinde, örneklem ölçümü arttıkça bazı algoritmaların performansının değiştiği gözlenmiştir. Bu değişim Lasso regresyonunda, örneklemin büyümesiyle birlikte *HKO*'da artış ve model açıklayıcılığında düşüş olarak ortaya çıkmıştır. Ridge regresyonunda, örneklem ölçümündeki artış, *HKO* ve R^2 üzerinde belirgin bir etki yaratmamıştır. Boosting algoritmaları açısından bakıldığında, LightGBM'in yeniden örnekleme tekniği kullanılıp kullanılmamasından bağımsız olarak en düşük *HKO* ve en yüksek R^2 değerlerine sahip algoritma olduğu gözlenmiştir. GradientBoost, çoğunluk sınıfın sayısının nadir sınıfın sayısına yaklaştırılması ile dağılımı dengeleyen Tomek bağlantısını kullandığında en iyi performansını elde etmiştir. Benzer şekilde, XGBoost ve AdaBoost için de bu durum geçerlidir. Doğru sınıflandırma performansını değerlendirmek için hesaplanan metriklerde ise farklı sonuçlar elde edilmiştir. Lasso regresyon, metrikler açısından en yüksek skorlarını örneklem ölçümünden bağımsız olarak SMOTE yeniden örneklemesinde elde etmiştir. Ridge regresyon, Prec ve Rec metrikleri açısından farklı örneklem ölçümleri ve yeniden örnekleme tekniklerinde iyi performans gösterirken, AUC açısından örneklem ölçümü ve yeniden örnekleme tekniğinden bağımsız en yüksek değerleri elde etmiştir. Boosting algoritmaları içinde Prec ve Rec açısından en iyi sonuçlar AdaBoost tarafından üretilmiştir. Tüm bu bulgular, algoritmaların farklı veri dengeleme tekniklerinde farklı performanslar sergilediğini göstermektedir. Kolay uygulanabilirliği sebebiyle dengesiz veri setlerinde Lasso veya Ridge regresyondan biri kullanılmak istenirse, Lasso regresyonunun tercih edilmesi durumunda performansın istikrarsız olabileceği, Ridge regresyonun ise daha stabil bir performans sergileyeceği göz önünde tutularak Ridge regresyon tercih edilebilir. Ancak sonuçlar genel olarak değerlendirildiğinde düşük hata ve yüksek açıklayıcılık oranları ile Boosting algoritmalarının Lasso ve Ridge regresyondan daha iyi performans sergilediği ve tercih edilebilirliklerinin daha fazla olduğu sonucu çıkarılabilir.

Teşekkür Yazar, değerli yorumları için Editörlere ve anonim hakemlere teşekkür eder.

Fon/Finansman Bilgileri Çalışma için herhangi bir mali sorumluluk yoktur.

Etik Kurul Onayı ve İzinler Çalışma etik kurul izni veya herhangi bir özel izin gerektirmemektedir.

Çıkar Çatışmaları/Çatışan Çıkarlar- Makale için herhangi bir çıkar çatışması yoktur.

Yazarların Katkısı Yazar makalenin son halini okumuş ve onaylamıştır.

Kaynaklar

[1] Bayman, O. E., & Dexter, F. (2021). Multicollinearity in logistic regression models. *Anesthesia & Analgesia*, *133*(2), 362-365. https://doi: 10.1213/ane.0000000000005593

[2] King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis, 9*(2), 137-163. https://doi:10.1093/oxfordjournals.pan.a004868

[3] Maalouf, M., & Trafalis, T. B. (2011). Robust weighted kernel logistic regression in imbalanced and rare events data. *Computational Statistics & Data Analysis*, *55*(1), 168-183. https://doi:10.1016/j.csda.2010.06.014

[4] Shrivastava, S., Jeyanthi, P. M., & Singh, S. (2020). Failure prediction of Indian Banks using SMOTE, Lasso regression, bagging and Boosting. *Cogent Economics & Finance*, *8*(1), 1729569. https://doi.org/10.1080/23322039.2020.1729569

[5] Rochayani, M. Y., Sa'adah, U., & Astuti, A. B. (2020). Finding biomarkers from a high-dimensional imbalanced dataset using the hybrid method of random undersampling and lasso. *Comtech: Computer, Mathematics and Engineering Applications*, *11*(2), 75-81. https://doi:10.21512/comtech.v11i2.6452

[6] Cahyana, N., Khomsah, S., & Aribowo, A. S. (2019). *Improving imbalanced dataset classification using oversampling and gradient Boosting* [Bildiri sunumu]. 5th international conference on science in information technology (ICSITech), China.

[7] Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data*, *7*, 1-47. https://doi.org/10.1186/s40537-020-00349-y

[8] Ashraf, M. T., Dey, K., & Mishra, S. (2023). Identification of high-risk roadway segments for wrong-way driving crash using rare event modeling and data augmentation techniques. *Accident Analysis & Prevention*, *181*, 106933. https://doi.org/10.1016/j.aap.2022.106933

[9] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267-288.

[10] Göv, A., & Kapkara Kaya, S. (2023). Türkiye örneğinde çevresel kalitenin belirleyicileri: lasso yaklaşımı. *Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, (54), 25-37. https://doi:10.30794/pausbed.1097352

[11] Yüzbaşı, B., & Pala, M. (2022). Ridge regresyon parametre seçimi: Türkiye'nin doğrudan yabancı yatırım örneği. *İstatistikçiler Dergisi: İstatistik ve Aktüerya*, *15*(1), 1-18.

[12] Mahesh, B. (2020). Machine learning algorithms - a review. *International Journal of Science and Research (IJSR)*, *9*(1), 381-386. https://doi: 10.21275/ART20203995

[13] Friedman, J. H. (2002). Stochastic gradient Boosting. *Computational Statistics & Data Analysis*, *38*(4), 367-378.

[14] Ali, Z. A., Abduljabbar, Z. H., Taher, H. A., Sallow, A. B., & Almufti, S. M. (2023). Exploring the power of extreme gradient Boosting algorithm in machine learning: A review. *Academic Journal of Nawroz University*, *12*(2), 320-334.

[15] Tyralis, H., & Papacharalampous, G. (2021). Boosting algorithms in energy research: A systematic review. *Neural Computing and Applications*, *33*(21), 14101-14117. https://doi.org/10.1007/s00521-021-05995-8

[16] Li, S., & Zhang, X. (2020). Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm. *Neural Computing and Applications*, *32*(7), 1971-1979. https://doi.org/10.1007/s00521-019-04378-4

[17] Wang, D. N., Li, L., & Zhao, D. (2022). Corporate finance risk prediction based on LightGBM. *Information Sciences*, *602*, 259-268. https://doi.org/10.1016/j.ins.2022.04.058

[18] Gu, Q., Sun, W., Li, X., Jiang, S., & Tian, J. (2023). A new ensemble classification approach based on Rotation Forest and LightGBM. *Neural Computing and Applications*, *35*(15), 11287-11308. https://doi.org/10.1007/s00521-023-08297-3

[19] Ying, C., Qi-Guang, M., Jia-Chen, L., ve Lin, G. (2013). Advance and prospects of AdaBoost algorithm. *Acta Automatica Sinica*, *39*(6), 745-758. https://doi.org/10.1016/S1874-1029(13)60052-X

[20] Hoens, T. R., & Chawla, N. V. (2013). Imbalanced datasets: from sampling to classifiers. H. He & Y. Ma (Ed.), *Imbalanced learning: Foundations, algorithms, and applications*, (s.43-59). Wiley Online Library.

[21] Wang, J., Xu, M., Wang, H., & Zhang, J. (2006). *Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding* [Bildiri sunumu]. In 8th international conference on signal processing (IEEE), China.

[22] Birla, S., Kohli, K., & Dutta, A. (2016). *Machine learning on imbalanced data in credit risk* [Bildiri sunumu]. In 2016 IEEE 7th annual information technology, electronics and mobile communication conference (IEMCON), Canada.

[23] Wang, Z. H. E., Wu, C., Zheng, K., Niu, X., & Wang, X. (2019). SMOTETomek-based resampling for personality recognition. *IEEE access*, *7*, 129678-129689. https://doi:10.1109/ACCESS.2019.2940061

[24] Werner de Vargas, V., Schneider Aranda, J. A., dos Santos Costa, R., da Silva Pereira, P. R., & Victória Barbosa, J. L. (2023). Imbalanced data preprocessing techniques for machine learning: a systematic mapping study. *Knowledge and Information Systems*, *65*(1), 31-57. https://doi.org/10.1007/s10115-022-01772-8

[25] Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315-1316. https://doi.org/10.1097/JTO.0b013e3181ec173d

[26] Keçeoğlu, Ç. R., Gelbal, S., & Doğan, N. (2016). Roc eğrisi yöntemi ile kesme puaninin belirlenmesi. *The Journal of Academic Social Science Studies, 50*(2), 553-562. http://dx.doi.org/10.9761/JASSS3564

[27] Oommen, T., Baise, L. G., & Vogel, R. M. (2011). Sampling bias and class imbalance in maximum-likelihood logistic regression. *Mathematical Geosciences*, *43*, 99-120. https://doi10.1007/s11004-010-9311-8

[28] Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, *17*(3), 299-310.

[29] LemaÃŽtre, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, *18*(17), 1-5.