NÖHÜ Müh. Bilim. Derg. / NOHU J. Eng. Sci., 2025; 14(2), 668-679

Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi
Niğde Ömer Halisdemir University Journal of Engineering Sciences

Araştırma makalesi / Research article

www.dergipark.org.tr/tr/pub/ngumuh / www.dergipark.org.tr/en/pub/ngumuh

CatBoost algoritmasının taşınmaz değerlemede kullanımı: Bayesian hiperparametre optimizasyonu ile karşılaştırmalı analiz

The use of Catboost algorithm in real estate valuation: A comparative analysis with Bayesian hyperparameter optimization

Berra Nur Tunç[1*], Ümit Haluk Atasever[2]

[1,2] *Erciyes Üniversitesi, Harita Mühendisliği Bölümü, 38030, Kayseri Türkiye*

Öz

Taşınmaz değerlemede, makine öğrenimi modelleri kullanılarak objektif, bilimsel ve hızlı tahminler elde edilmektedir. Bu çalışmada, hiperparametre optimizasyonu yapılmış farklı makine öğrenimi modelleri kullanılarak taşınmaz değerlemede en tutarlı ve başarılı sonucu veren model belirlenmiştir. Özellikle CatBoost regresyonu, modern makine öğrenimi ihtiyaçlarına uygun olarak geliştirilmiş, yüksek doğruluk ve hız sunan bir model olarak ön plana çıkmaktadır. Çalışmada CatBoost'un yanı sıra Destek Vektör Regresyonu, Lasso Regresyonu, Karar Ağaçları Regresyonu ve AdaBoost Regresyonu da değerlendirilmiştir. Deneysel hesaplamalar için Boston şehrinde toplanmış 506 konutun çeşitli öznitelikleri ve fiyatlarına sahip bir veri seti kullanılmıştır. Hata metrikleri karşılaştırıldığında, optimize edilmiş CatBoost regresyonu, tüm modeller arasında en yüksek performansı göstermiştir. Özellikle, literatürdeki diğer yöntemlere kıyasla daha başarılı tahminler sunarak, taşınmaz değerleme çalışmalarında öne çıkmıştır. Destek Vektör Regresyonu ise nispeten daha düşük başarı sergilemiştir.

Anahtar kelimeler: Taşınmaz Değerleme, Makine Öğrenimi Modelleri, CatBoost Regresyonu, Hiperparametre Optimizasyonu

Abstract

In real estate valuation, machine learning models enable objective, scientific, and rapid predictions. This study aimed to determine the most consistent and accurate model for real estate valuation using various machine learning models with hyperparameter optimization. Specifically, CatBoost regression stands out as a modern model developed to meet contemporary machine learning needs, offering high accuracy and speed. In addition to CatBoost, Support Vector Regression, Lasso Regression, Decision Tree Regression, and AdaBoost Regression were also evaluated. For experimental calculations, a dataset comprising various attributes and prices of 506 properties collected in Boston was used. When error metrics were compared, the optimized CatBoost regression demonstrated the highest performance among all models. Particularly, it provided more accurate predictions than other methods in the literature, establishing itself as a standout approach in real estate valuation studies. Support Vector Regression, on the other hand, exhibited relatively lower success.

Keywords: Real Estate Valuation, Machine Learning Models, CatBoost Regression, Hyperparameter Optimization

1 Giriş

Taşınmaz, arazi, bina, bahçe gibi yer değiştirilmesi mümkün olmayan malları ifade eden bir kavramdır. Taşınmaz değerlemesi, bir taşınmazın konumu, fiziksel özellikleri ve çevresel faktörlerini dikkate alarak değerini belirlemek amacıyla yapılan bilimsel değerlendirme sürecidir. Bir taşınmazın tam değerini belirlemek gerçekte mümkün değildir. Çünkü her taşınmaz, bulunduğu konum ve kullanım şekline göre farklı özellikler sergiler ve bu özellikler, kişiden kişiye önem olarak değişmektedir. Bu nedenle taşınmaz değerlemesinde kriterlerin önem derecesi olarak kesin bir bilgi yoktur. Ancak, bilimsel yöntemler kullanılarak yapılan değerleme işlemleri ile taşınmazın değeri daha doğru bir şekilde belirlenebilmektedir [1].

Bir taşınmazın değeri farklı fiziksel ve çevresel özelliklerin bir araya gelmesiyle belirlenmektedir. Bu yüzden taşınmazın değerini doğru şekilde yansıtacak özelliklerin belirlenmesi ve analiz edilmesi büyük önem taşımaktadır. Özelliklerin seçimi yapılırken, kriterler arasındaki ilişkilerin modelin doğruluğu üzerindeki etkisi dikkate alınmalıdır [2].

Değerleme sürecinde bir taşınmazın piyasa değerini belirleyen unsurlar; yapısal özellikleri, bulunduğu konum, kullanım alanı, ulaşım imkanları gibi çevresel faktörler, toplumsal yapı gibi sosyal faktörler ve kişisel etmenler olarak sıralanabilmektedir [3]. Taşınmaz değerleme sürecinde kullanılan parametrelerin doğru, tutarlı ve tarafsız bir şekilde değerlendirilmesi amacıyla çeşitli yöntemler geliştirilmiştir. Bu yöntemler arasında klasik, istatistiksel analizler, Coğrafi Bilgi Sistemleri (CBS), çok kriterli karar analizleri ve makine öğrenimi teknikleri sayılmaktadır. Özellikle son yıllarda birçok akademik çalışmada kullanılan makine öğrenimi yöntemleri çok sayıda taşınmaz verisini bir arada inceleyerek bilimsel bir yaklaşımla değerleme süreçlerini sağlamaktadır [4, 5]. Son yıllarda, algoritmaların performansını artırmak amacıyla modelin eğitildiği

* Sorumlu yazar / Corresponding author, e-posta / e-mail: berranursen@erciyes.edu.tr (B. N. Tunç)
Geliş / Received: 06.12.2024 Kabul / Accepted: 14.03.2025 Yayımlanma / Published: 15.04.2025
doi: 10.28948/ngumuh.1597288

hiperparametrelerin ayarlanması veya optimize edilmesi amacı ile hiperparametre optimizasyonu üzerine çalışmalar yapılmaktadır. Hiperparametreler, makine öğrenimi modellerinin yapılandırılması sürecinde kullanıcı tarafından belirlenen ve modelin öğrenme sürecine rehberlik eden parametrelerdir [6]. Hiperparametre optimizasyonu makine öğrenimi modellerinin sürecini direkt olarak etkileyen ve değerleri en üst düzeye çıkarabilen hiperparametre kombinasyonlarının bulunması sürecidir. Bu süreç çoğunlukla deneme yanılma, grid araması, rastgele arama ve Bayesian hiperparametre optimizasyonu gibi farklı teknikler aracılığıyla uygulanır [6, 7].

Literatürde makine öğrenimi modelleri kullanılarak farklı birçok çalışma gerçekleştirilmiştir.

Türkan vd. [4] çalışmalarında kent ölçeğinde toplu taşınmaz değerlemesi sürecinde farklı makine öğrenimi tekniklerinin uygulanmasını ele almaktadır. Çalışmanın amacı, Niğde kentinde 1200 taşınmaza ait verileri kullanarak Lineer Regresyon, Yapay Sinir Ağları, Regresyon Ağaçları, Destek Vektör Regresyonu ve Gaussian Process Regresyonu gibi algoritmalarla taşınmaz değer tahmini gerçekleştirmektir. Çalışmaları sonucunda, eğitim verileri üzerinde en başarılı modelin Gaussian Process Regresyonu olduğu, ancak test verilerinde en yüksek doğruluğa Yapay Sinir Ağları ile ulaşıldığı belirlenmiştir. Çalışma, kent ölçeğinde toplu taşınmaz değerlemede makine öğrenimi tekniklerinin etkinliğini ortaya koyması ve Yapay Sinir Ağları'nın en güvenilir tahminleri sağladığını göstermesi açısından literatüre önemli bir katkı sunmaktadır.

Hazer vd. [6] çalışmalarında küçük ölçekli kentlerde taşınmaz değerlemesi sürecinde hiper-optimize makine öğrenimi tekniklerinin uygulanmasını ele almaktadır. Çalışmanın amacı, 2022 ve 2023 yıllarına ait taşınmaz verilerini kullanarak Bayes Tekniği ile optimize edilmiş regresyon modellerini kullanarak değer tahmini gerçekleştirmektir. Çalışmaları eğitim verileri üzerinde en başarılı modelin Çekirdek Regresyonu, test verilerinde ise Topluluk Regresyonu olduğu sonucuna varmışlardır. Çalışma, toplu taşınmaz değerlemede hiper-optimize edilmiş makine öğrenimi tekniklerinin heterojen özelliklere sahip alanlarda bile güvenilir sonuçlar sağlayabileceğini kanıtlaması açısından literatüre önemli bir katkı sunmaktadır.

Baur vd. [8] çalışmalarında konut fiyatlarının tahmininde yapı özelliklerinin haricinde ilan metnini de dikkate alarak konut değerleme modellerinin doğruluğunu artırmayı amaçlamışlardır. Berlin'deki kiralık daireler ve Los Angeles'taki ev satış teklifleri verilerini kullanarak çeşitli makine öğrenimi modellerini değerlendirmişlerdir. Sonuç olarak ilan metninin modele dahil edilmesinin ortalama mutlak hata (MAE) oranını azalttığını belirlemişlerdir.

Hernes vd. [9] çalışmalarında 2016-2022 yıllarında Wroclaw'daki arsa emlak verilerini ele alıp regresyon modellerinden rastgele orman ve yapay sinir ağlarının tahmin performansını değerlendirerek taşınmaz değerleme sürecinde makine öğrenimi modellerinin uygulanabilirliğini araştırmayı amaçlamışlardır. Çalışmalarının sonucunda en düşük tahmin hatası oranı olan %13 MAPE ile yapay sinir ağı modelinin en yüksek doğruluk seviyesine ulaştığını belirlemişlerdir. Bu sonuçlar ile makine öğrenimi modellerinin taşınmaz değerleme için etkili sonuçlara ulaşılabileceğini ortaya koymuşlardır.

Jafary vd. [10] çalışmalarında Melbourne Metropol Bölgesi'nde fiziksel, coğrafi, sosyoekonomik ve çevresel faktörlere dayalı olarak XGBoost, Destek Vektör Regresyonu, Rastgele Orman ve Derin Sinir Ağı olmak üzere dört farklı makine öğrenimi modelini kullanarak otomatik değerleme modellerinin performansını karşılaştırmayı amaçlamışlardır. Sonuçlara göre XGBoost diğer modellere göre üstün performans gösterdiğini belirlemişler.

Konhauser vd. [11] çalışmalarında konut binalarının enerji verimliliği özelliği ile mülk değeri arasındaki ilişkiyi analiz etmeyi amaçlamışlardır. Birleşik Krallıktaki Enerji Performans Sertifikaları ve emlak işlem verilerini birleştirerek XGBoost ve CatBoost gibi iki gelişmiş makine öğrenimi modelini uygulamışlardır. Enerji verimli binaların Londra dışındaki bölgelerde finansal fayda sağladığı sonucuna ulaşmışlardır. Çalışmadan elde ettikleri sonuçlar ile enerji verimliliği yatırımlarının ekonomik faydalarını ve karbonsuz bir topluma geçişteki önemi vurgulamışlardır.

Grzybowska vd. [12] çalışmalarında Orta Avrupa'daki dört ülkede konut kalabalıklığını ve sosyoekonomik faktörleri incelemeyi amaçlamışlardır. Araştırmalarında 2022 yılına ait gelir ve yaşam koşulları üzerine en son istatistiksel verileri kullanmışlardır. Makine öğrenimi modeli olarak Random Forests, Balanced Random Forests, Extreme Gradient Boosting ve yeni bir yöntem olan CatBoost gibi modelleri uygulamışlardır. CatBoost modelinin diğer modellere kıyasla doğruluk ve eğri altındaki alan (AUC) açısından en yüksek performansı sergilediğini belirlemişlerdir. Çalışmaları sonucunda, ülke faktörünün konut yoksulluğunun belirlenmesinde önemli bir etkiye sahip olduğunu vurgulamışlardır.

Kalliola vd. [13] tarafından gerçekleştirilen çalışmada, Helsinki'deki taşınmaz fiyatlarını tahmin etmek için yapay sinir ağlarının (YSA) hiperparametre optimizasyonunu ele almaktadır. Çalışmalarında YSA modellerinin hiperparametrelerini Bayesian optimizasyonu kullanarak en iyi hale getirmek ve fiyat tahmin doğruluğunu artırmayı amaçlamışlardır. Çalışma sonucunda hiperparametre optimizasyonunun model performansında önemli bir iyileşme sağlandığını göstermişlerdir. Bu çalışma, taşınmaz değerleme sürecinde makine öğrenimi modellerinin doğruluğunu artırmada Bayesian optimizasyonunun önemini ortaya koymaktadır.

Aydınoğlu vd. [14] çalışmalarında, toplu taşınmaz değerleme süreçlerinde farklı makine öğrenme algoritmalarının etkinliğini ve konumsal/konumsal olmayan özniteliklerin tahmin doğruluğuna etkilerini incelemektedir. Çalışmanın amacı, İstanbul Pendik ilçesindeki 1475 taşınmaza ait verileri kullanarak Çoklu Doğrusal Regresyon (ÇDR), Genelleştirilmiş Doğrusal Model (GDM), Destek Vektör Makineleri (DVM), Karar Ağaçları (KA) ve Rastgele Orman (RO) algoritmaları ile taşınmaz değer tahmini gerçekleştirmektir. Çalışmalarında, en yüksek doğruluğa Rastgele Orman algoritması ile ulaşıldığını, en düşük doğruluğun ise Karar Ağaçları ve GDM modellerinde gözlendiğini belirlemişlerdir. Ayrıca, konumsal ve konumsal

olmayan veri setlerinin model doğruluğuna etkileri incelendiğinde, konumsal verilerin taşınmaz değer tahmininde daha büyük bir etkiye sahip olduğu tespit edilmiştir. Çalışma, toplu taşınmaz değerlemede makine öğrenme algoritmalarının farklı veri setleriyle test edilerek etkinliğinin karşılaştırılması açısından literatüre önemli bir katkı sunmaktadır.

Literatürde taşınmaz değerleme süreçlerinde makine öğrenimi modellerinin kullanımı yaygınlaşmış olsa da hiperparametre optimizasyonu ile iyileştirilmiş modellerin karşılaştırmalı analizi çalışmaları sınırlıdır. Bu çalışma, Bayesian hiperparametre optimizasyonu ile iyileştirilmiş modern makine öğrenimi yöntemlerinden biri olan ve kategorik değişkenlerle doğrudan çalışabilme avantajına sahip CatBoost regresyon yönteminin ve literatürde yaygın olan diğer regresyon yöntemlerini kullanılarak, taşınmaz değerleme de en başarılı sonucu üreten modeli elde etmeyi amaçlamaktadır. Çalışma taşınmaz değerleme alanında literatürde kullanılan geleneksel makine öğrenimi yöntemlerinden farklı olarak modern yöntemlerden biri olan CatBoost algoritmasında hiperparametre optimizasyonu kullanımı ile literatürdeki diğer çalışmalardan farklılaşmaktadır.

Çalışmada Boston veri seti [15] kullanılmış olup makine öğrenimi yöntemlerinden CatBoost Algoritması, Destek Vektör Regresyonu (DVR), Lasso Regresyonu ve Karar Ağaçları Regresyonu, AdaBoost Regresyonu modelleri kullanılmıştır. Çalışmada hiperparametrelerinin doğru belirlenebilmesi için hiperparametre optimizasyon yöntemlerinden Bayesian hiperparametre optimizasyonu ve farklı regresyon modellerinin hata metrikleri kullanılarak bu modellerin birbirine göre kıyaslaması gerçekleştirilmiştir. Hiperparametre optimizasyonu ile iyileştirilmiş yöntemlerin hem eğitim hem test verileri için genel olarak sonuçlarına bakıldığında yöntemlerin kendi içinde tutarlı sonuçlar elde ettiği görülmüştür. Ancak yöntemleri kıyaslandığında CatBoost regresyon yöntemi en başarılı sonucu verirken DVR regresyon yöntemi ise görece daha başarısız sonucu vermiştir.

2 Materyal ve metot

Bu çalışma, hiperparametre optimizasyon yöntemlerinden Bayesian hiperparametre optimizasyonu ile iyileştirilmiş regresyon yöntemleri kullanılarak, taşınmaz değerleme de en tutarlı ve başarılı sonucu üreten modeli elde etmeyi amaçlamaktadır. Çalışmada daha önce birçok uygulamada kullanılmış olan Boston veri seti kullanılmıştır. Bu veri seti Massachusetts eyaletinin başkenti olan Boston şehrine toplam 506 konuta ait verileri içermektedir. Bu konutlara ait toplam 13 özellik verisi bulunmaktadır. Veriler tamamen rastgele bir şekilde %70'i eğitim ve %30'u test verisi olmak üzere bölünmüştür. Boston Veri setindeki 506 adet konut için seçilen 13 adet özellik Tablo 1' de verilmiştir.

Kullanılan bu veri setinde konum bilgisiolarak. Boston'daki beş ana iş merkezine olan mesafe ağırlığı, otoyollara erişim indeksi, Charles Nehrine yakınlık gibi konumsal faktörler bulunmaktadır. Bu veri setinin seçilme nedeni taşınmaz fiyatlarını etkileyen geniş bir değişken skalasına sahip olması ve literatürde yaygın olarak karşılaştırmalı analizler için kullanılmasıdır. Özellikle fiziksel özellikler (oda sayısı, bina yaşı), ekonomik özellikler (emlak vergisi, düşük gelirli nüfus yüzdesi) ve konum özellikleri (otoyollara erişim, iş merkezlerine uzaklık) gibi değişkenlerin taşınmaz fiyatları üzerindeki etkisinin önemidir. Çalışmada bu değişkenler hiperparametre optimizasyonu uygulanarak makine öğrenimi modellerinde kullanılmış ve taşınmaz değerleme süreçlerindeki etkileri kapsamlı bir şekilde analiz edilmiştir.

Tablo 1. Boston veri setindeki özellikler

Özellik	Açıklama
0	Crim: Kişi başına düşen suç oranı (kasaba bazında)
1	Zn: 2350 m²'den büyük arsalara sahip konut alanının oranı
2	Indus: Kasaba başına düşen ticaret dışı iş alanlarının oranı
3	Chas: Charles Nehri için değişken (1: nehre sınırsa; 0: değilse)
4	Nox: Azot oksit konsantrasyonu (milyonda birim)
5	Rm: Konut başına ortalama oda sayısı
6	Age: 1940'tan önce inşa edilmiş, sahibi tarafından kullanılan birimlerin oranı
7	Dis: Boston'daki beş ana iş merkezine olan mesafe ağırlığı
8	Rad: Radyal otoyollara erişim indeksi
9	Tax: Her 10.000 dolar başına tam değerli emlak vergisi
10	Ptratio: Kasabadaki öğrenci-öğretmen oranı
11	B: kasabadaki siyahi nüfus oranını içeren bir ölçüt
12	Lstat: Düşük statüye sahip nüfus yüzdesi

2.1 Makine öğrenimi modelleri

2.1.1 CatBoost algoritması

CatBoost regresyon modeli Yandex araştırmacıları tarafından karar ağacı gradyanını arttırmak amacıyla geliştirilmiş ve kategorik verileri ön işlemeye ihtiyaç duymaksızın doğrudan veriyi kullanabilen etkili bir öğrenme modelidir [16]. CatBoost modelinin büyük eğitim verisi gereksinimlerine bakılmaksızın verimli sonuçlar elde edebilmesi, modeli diğer modellerden ayıran en belirgin özelliktir. CatBoost'un gerçek dünya uygulamalarında etkili sonuçlar elde etmesini sağlayan avantajlar arasında kategorik değişkenleri otomatik olarak etkili bir şekilde işleyebilmesi, sınıflandırma doğruluğunun yüksek olması, hiperparametrelerin kolaylıkla ayarlanabilmesi için optimizasyon mekanizması içermesi ve aşırı öğrenmenin kontrol edilerek azaltılması olarak sıralanabilmektedir [16].

CatBoost yönteminin temelinde zayıf regresyon modellerinin bir araya getirilerek daha güçlü bir model oluşturulabileceği fikri yatmaktadır. $L(.)$ kayıp fonksiyonudur ve model beklenen kayıp fonksiyonu gradyan inişi yöntemi ile Denklem (1)'deki gibi minimize ederek eğitilir [17]:

$$h^t = \arg\min_{h \in H} L(F^{t-1} + h) = \arg\min_{h \in H} \mathbb{E}L\left(y, F^{t-1}(x) + h(x)\right) \quad (1)$$

Burada *y* çıktıyı temsil eder ve *h*, *H* fonksiyon ailesinden seçilen bir gradyan adım fonksiyonudur. Bu fonksiyonun hesabı Denklem (2)'deki gibidir;

$$h(x) = \sum_J b_j I\{x \in R_j\} \quad (2)$$

Bu denklemde R_j, karar ağacının yapraklarına karşılık gelen bölgeleri, b_j ise karşılık gelen bu bölgelerin tahmin değerlerini ifade eder. I ise bir gösterge fonksiyonudur [16, 18, 19].

2.1.2 *Destek vektör regresyonu*

Destek vektör regresyonu Vapnik ve Cortes [20] tarafından geliştirilen ve özellikle verilerin lineer olarak ayrılamaması durumunda iyi performans gösteren bir makine öğrenme modelidir [4]. DVR temel olarak bir aşırı düzlem oluşturarak çalışmaktadır. Bu düzlem, veri noktaları ile bu noktaların tahmin edilen regresyon değerleri arasındaki farkı kontrol etmeye odaklanmaktadır. Amaç, bu farkın belirli tolerans aralığı içinde kalmasını sağlamaktır. Bu şekilde model, tahmin edilen değerlerle gerçek değerler arasındaki uyumu optimize etmektedir [6, 21].

DVR regresyon fonksiyonun matematiksel ifadesi Denklem (3)'deki gibidir.

$$f(x) = \sum_{i=1}^{n} (w_i, \varphi(x_i)) + b \quad (3)$$

Burada *f(x)* tahmin edilen değeri, w_i modelin öğrenme sürecinde her bir destek vektörüne ne kadar önem verileceğini belirler, $\varphi(x_i)$ destek vektörlerini tanımlar ve verileri daha yüksek boyutlu bir uzaya dönüştürmek için kullanılan bir fonksiyon, *b* ise aşırı düzlemin veri noktalarına göre konumunu ayarlayan bir sabittir [6].

2.1.3 *Lasso regresyonu*

Lasso regresyonu, verilerin belirli hata sınırları içerisinde kalacak şekilde hedef değişken için tahmin performansını artırmayı amaçlamaktadır. Lasso regresyonda L_1 düzeltmesi kullanılmaktadır. L_1 düzeltmesi $\sum_{j=1}^{p}|\beta_j|$ olarak ifade edilmektedir [22]. Lasso regresyonunun matematiksel ifadesi Denklem (4)'deki gibidir.

$$\hat{\beta}_{Lasso} = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = AKT + \lambda\sum_{j=1}^{p}|\beta_j| \quad (4)$$

Burada λ ceza terimini temsil eder ve sıfıra yaklaştıkça Lasso doğrusal regresyona yakın sonuç elde etmektedir. Bu yüzden λ parametresinin seçimi modelin temsili için oldukça önemlidir. y_i gerçek değerleri, β_0 model sabit terimini, β_j *j*'inci bağımsız değişkenin regresyon katsayısını, x_{ij} *i*'inci gözlemin *j*'inci bağımsız değişkenine karşılık gelen değeri, *AKT* artık kareler toplamını ifade etmektedir[22].

2.1.4 *Karar ağaçları regresyonu*

Karar ağaçları regresyonu, tahmini gerçekleştirilen değişken değerinin bağımsız değişkenler yardımı ile tahmin edilmesinde kullanılan bir makine öğrenmesi yöntemdir. Karar ağaçlarının temel amacı değer kümesini $(x_1, x_2, \ldots, x_p)$ j sayıda yaprağa ayırarak $(R_1, R_2, \ldots, R_j)$ her yaprak için tahminlerde bulunmaktır. Regresyon ağaçlarında, değer noktaları yapraklara ayrıldıkça her yaprak için gözlemlenen sonuçların ortalaması hesaplanmaktadır. R_j yaprağındaki tahmin Denklem (5)'deki şekildedir.

$$\hat{y}_{R_j} = \frac{1}{n_j}\sum_{i \in R_j} y_i \quad (5)$$

Burada $\hat{y}_{R_j}$ tahmin edilen değeri, n_j R_j yaprağına düşen veri sayısı, y_i R_j yaprağındaki her bir gözlemin gerçek değerini, R_j karar ağacının *j*'inci yaprağını ifade etmektedir [22].

2.1.5 *AdaBoost regresyonu*

AdaBoost regresyonu, örneklerin özellik uzayını yinelemeli bir şekilde tarayıp eğitim verilerinin ağırlıklarını belirlemektedir. Bu regresyonun asıl amacı zayıf modellerin güçlü yönlerini alarak en iyi modeli oluşturmaktır. AdaBoost, birden fazla model üretir ve her modelin avantajlı yanlarını bir araya getirerek bu modellerden başarılı bir model çıkarmaya çalışmaktadır. AdaBoost algoritması, zayıf modelleri birleştirerek en başarılı modelin sonucunu Denklem (6) ile üretmektedir.

$$h_f(x) = \inf[y \in Y : \sum_{t:h_{t(x)} \le y} \log\left(\frac{1}{\alpha_t}\right) \ge \frac{1}{2}\sum_t \log\left(\frac{1}{\alpha_t}\right)] \quad (6)$$

Burada $h_f(x)$ tahmin edilen değeri $h_{t(x)}$ zayıf öğrenicileri, α_t model ağırlığını ifade etmektedir [23, 24].

2.2 *Çalışmada kullanılan hata metrikleri*

Makine öğrenim modelleri ile veri seti ile ilgili yapılan tahminlerin ne kadar başarılı olduklarının belirlenebilmesi için hata metrikleri kullanılmaktadır. Bu hata metriklerinden, MSE (Karesel Ortalama Hata), MAE (Ortalama Mutlak Hata), R^2 (R-Kare), MAPE (Ortalama Mutlak Yüzdesel Hata), COD (Tahminlerin Gerçek Değerler ile Sapma Oranı) ve PRD (Fiyata Bağlı Fark Oranı) kullanılmıştır.

Bu hata metriklerine ait formüller sırasıyla Denklem (7), Denklem (8), Denklem (9), Denklem (10), Denklem (11), Denklem (12), Denklem (13) ile ifade edilmiştir. Özellikle COD ve PRD gayrimenkul değerleme amacı ile kullanılan bir ölçüttür. COD, tahmin edilen değerlerin standart sapmasının gerçek değere oranını temsil eden dağılım katsayısı iken PRD, bir gayrimenkulün piyasa değeri ile tahmin edilen değeri arasındaki tutarlılığı ölçmek için kullanılmaktadır [5].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (7)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (8)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (9)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|} \quad (10)$$

$$COD = \frac{100}{m} * (\frac{\sum_{i=1}^{n}\left|\frac{y_i}{\hat{y}_i} - m\right|}{n-1}) \quad (11)$$

$$m = medyan(y_i * \sqrt{\hat{y}_i}) \quad (12)$$

$$PRD = \frac{mean(\frac{y_i}{\hat{y}_i})}{\sum_{i=1}^{n} y_i / \sum_{i=1}^{n} \hat{y}_i} \quad (13)$$

Burada y_i gerçek değeri, $\hat{y}_i$ tahmin edilen değeri ve $\bar{y}_i$ ise gerçek değerlerin ortalamasını ifade etmektedir.

2.3 Bayesian hiperparametre optimizasyonu

Bayesian hiperparametre optimizasyonu, makine öğrenimi modellerinin hiperparametrelerini sistematik olarak optimize etmek için kullanılan bir yöntemdir. Geleneksel ızgara arama veya rastgele arama yöntemlerinden farklı olarak, Bayesian optimizasyon olasılıksal modeller kullanarak hiperparametrelerin performansını tahmin eder. Bayesian optimizasyon hiperparametrelerin model performansı üzerindeki etkisini tahmin etmek için bir vekil model (örneğin Gaussian Süreci, Rastgele Orman veya Ağaç Yapılı Parzen Tahmincisi) kullanır. Daha sonra yeni hiperparametre değerleri seçmek için bir kazanım fonksiyonu devreye girer. Süreç yeni hiperparametre ve hata değerleri kullanılarak vekil modelin sürekli güncellenmesiyle iteratif olarak devam eder [25]. Bayesian optimizasyon ilk aşamasında hiperparametre arama alanı tanımlanır. Buradaki amaç optimize edilecek hiperparametrelerin türü (örneğin sürekli, kategorik veya tamsayı) ve aralıkların belirlenmesidir. Daha sonra başlangıçta birkaç rastgele hiperparametre kombinasyonu seçilerek model üzerinde değerlendirilerek hata hesaplanır ve bu sonuçlarla vekil model oluşturulur. Bu vekil model ile kazanım fonksiyonunun en iyi performansı vereceği tahmini hiperparametreleri önermesini sağlanır. Bu hiperparametreler model üzerinde değerlendirilerek sonuçlar vekil modele eklenir ve süreç iteratif olarak tekrarlanır [25].

Bayesian optimizasyonun en büyük avantajı, modelin performansını tahmin etmek için daha az sayıda denemeye ihtiyaç duymasıdır. Ayrıca arama alanında hem keşif (bilinmeyen bölgelerin araştırılması) hem de sömürü (bilinen iyi bölgelerin daha fazla değerlendirilmesi) arasında dengeli bir strateji izler. Bayesian optimizasyon hem sürekli hem de kategorik hiperparametreleri optimize edebilmesiyle nedeniyle hemen hemen tüm regresyon modellerinin hiperparametrelerini optimize edilmesi için kullanılabilir [26].

Literatürde yapılan önceki çalışmalarla kıyaslandığında bulguların büyük ölçüde örtüştüğü görülmektedir. Baur vd. [8] tarafından yapılan konut değerleme çalışmasında, makine öğrenimi modellerinde hiperparametre optimizasyonunun hata oranlarını düşürdüğü ifade edilmiş ve çalışmamızda da hiperparametre optimizasyonunun doğruluk üzerindeki olumlu etkisi gösterilmiştir. Hernes vd. [9] tarafından yapılan çalışmada da hiperparametre optimizasyonu uygulanmış modellerin daha yüksek doğruluk sağladığı belirlenmiştir. Jafary vd. [10] tarafından yapılan otomatik taşınmaz değerleme çalışmasında, XGBoost ve DVR'nin performansı karşılaştırılmış ve optimizasyon sonrası XGBoost'un en başarılı model olduğu belirlenmiştir. Bu çalışmada DVR başarısız bulunmuş olup, çalışmamızın bulguları ile örtüşmektedir. Konhauser vd. [11] tarafından yapılan çalışmada XGBoost ve CatBoost regresyonları karşılaştırılmış ve CatBoost'un yüksek doğruluk sunduğu görülmüştür. Kalliola vd. [13] tarafından yapılan çalışma hiperparametre optimizasyonunun taşınmaz değerleme sürecindeki önemini vurgulamıştır. Bu çalışmada kullanılan Bayesian optimizasyon yönteminin model performansını artırdığı görülmüştür. Prokhorenkova vd. [18] tarafından yapılan çalışmadaki CatBoost'un hiperparametre optimizasyonu sonrası üstün performans gösterdiği bulgularıyla paralellik göstermektedir.

3 Bulgular ve tartışma

Regresyon modellerinin hiperparametre optimizasyonunda model başarısını en üst düzeye çıkarmak için kontrol parametrelerinin hangisinin ayarlanacağının belirlenmesi büyük önem taşımaktadır [27]. Hiperparametre optimizasyonu uygulanmış regresyon modellerinde, veriler rastgele bir şekilde %70'i eğitim ve %30'u test verisi olacak şekilde eğitim ve test aşamaları için bölünmüştür. Bu bağlamda kullanılan regresyon modellerinde optimize edilen hiperparametreler literatürde yaygın olarak kullanılan değer aralıklarına dayandırılmış olup, Tablo 2'de bu aralıklar, Tablo 3'te ise optimizasyon sonucunda en iyi performansı gösteren hiperparametreler verilmiştir.

Şekil 1'de verilen korelasyon matrisi, çeşitli özelliklerin ve bir hedef değişken arasındaki korelasyonları göstermektedir. Korelasyon matris grafiği genellikle ısı haritası olarak görselleştirilmektedir. Korelasyon katsayılarının renklerle ifade edildiği ve böylece yüksek veya düşük korelasyonların kolayca gözlemlenebildiği bir yöntemdir. Özellik-8 (Radyal otoyollara erişim indeksi) ve özellik-9 (Her 10.000 dolar başına tam değerli emlak vergisi) hedefle pozitif korelasyonlar (sırasıyla 0.62 ve 0.58) sergilerken, özellik-11 (kasabadaki siyahi nüfus oranını içeren bir ölçüt) ise negatif bir korelasyona (-0.39) sahiptir. Ayrıca özellikler arasındaki korelasyonların genel olarak çok yüksek olmaması, regresyon süreçlerinde kullanılmalarını oldukça uygun hale getirmektedir.

Şekil 2-6' da sunulan grafikler kurulan modelin gerçek değerler ile tahmin değerleri arasındaki ilişkiyi

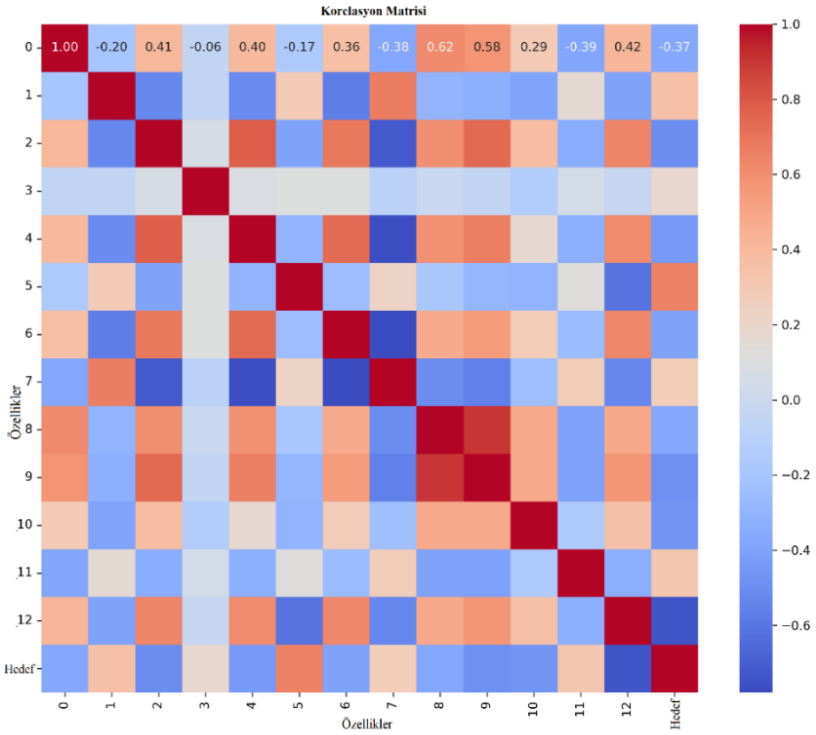
değerlendirmek amacı ile kullanılmaktadır. Grafik üzerindeki noktalar kesikli çizgi etrafında yoğunlaşmışsa model genellikle tutarlı tahminler yapmıştır aksi durumda yani noktalar geniş bir alana yayılmışsa modelin tahminlerinde hata yaptığı anlamına gelmektedir.

Tablo 2. Regresyon yöntemleri için hiperparametre optimizasyon arama limitleri [28, 29, 30, 31, 32]

Regresyon Modelleri	**Bayes Tekniği ile Optimize Edilmiş Hiperparametreler**
CatBoost Regresyonu	Öğrenme oranı (learning rate): [1e-5- 1e-1] Bagging sıcaklığı (bagging temperature): [0-1] Sınır sayısı (border count): [32-255] Derinlik (depth): [3-11] İterasyon sayısı (iteration): [50-200] L_2 yaprak düzenleme parametre değeri (L_2 leaf regularization): [1-10]
Destek Vektör Regresyonu	Ceza terimi parametresi (C): [1e-2- 1e2] Epsilon değeri: [1e-3- 1e-1] Kernel: [Lineer- RBF]
Lasso Regresyonu	Düzenleme katsayısı (alpha): [0.01- 10.00] Modelin eğitilmesi için maksimum iterasyon sayısı (max. iter): [100-1000] Tolerans (tol) parametresi: [1e-5- 1e-1]
Karar Ağaçları Regresyonu	Maksimum derinlik (max. depth): [3-21] Her bir yaprak için minimum örneklem sayısı (min. samples leaf): [1-10] Minimum örneklem bölme oranı (min. samples split): [0.1-1.0]
AdaBoost Regresyonu	Öğrenme oranı (learning rate): [1e-5- 1e-1] Temel öğrenici sayısı (n estimators): [50-200]

Tablo 3. Modeller için hesaplanmış en iyi hiperparametreler

Regresyon Modelleri	**Modeller için hesaplanmış en iyi hiperparametreler**
CatBoost Regresyonu	Öğrenme oranı (learning rate): 0.0955 Bagging sıcaklığı (bagging temperature): 0.0795 Sınır sayısı (border count): 54 Derinlik (depth): 5 İterasyon sayısı (iteration): 176 L2 yaprak düzenleme parametre değeri (L2 leaf regularization): 5.6554
Destek Vektör Regresyonu	Ceza terimi parametresi (C): 0.6593 Epsilon değeri: 0.00001 Kernel: Lineer
Lasso Regresyonu	Düzenleme katsayısı (alpha): 0.01 Modelin eğitilmesi için maksimum iterasyon sayısı (max. iter): 841 Tolerans (tol) parametresi: 0.0949
Karar Ağaçları Regresyonu	Maksimum derinlik (max. depth): 16 Her bir yaprak için minimum örneklem sayısı (min. samples leaf): 2 Minimum örneklem bölme oranı (min. samples split): 0.1
AdaBoost Regresyonu	Öğrenme oranı (learning rate): 0.0299 Temel öğrenici sayısı (n estimators): 200

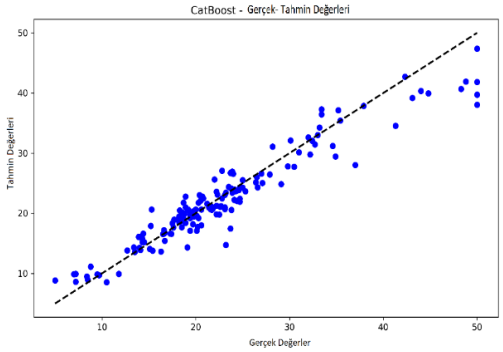


Şekil 1. Veri setine ait korelasyon matrisi

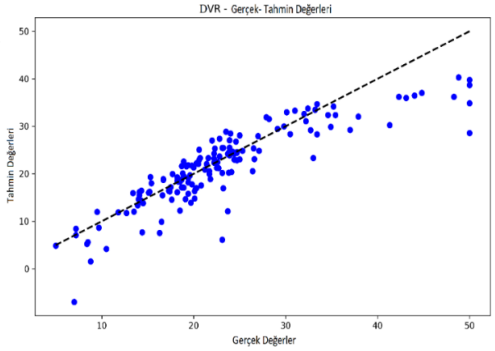
Eğitim verileri için gerçekleştirilen modeller incelendiğinde tümünde orta değer aralığında (yaklaşık 10-30) verilerin yoğunlaştığı ve bu değer aralığında noktaların çoğunun ideal çizgiye yakın yer aldığı gözlenmektedir. Bu da bu değer aralığında modellerin başarılı olduklarını göstermektedir. Ancak düşük ve yüksek değerlerde tahminlerin gerçek değerlerden sapmış oldukları gözlenmektedir. Bu da modelin bu değer aralıklarını tahmin etme konusunda zorlandığını ve bu bölgelerde hataların artabileceğini göstermektedir.

CatBoost, tahminlerin genel doğruluğu ve gerçek değerlere yakınlığı açısından en iyi performansı göstermiş ve noktaların çoğu ideal doğruya oldukça yakın yerleşmiştir. AdaBoost ve Karar Ağaçları modelleri orta düzeyde performans sergilemişlerdir. Her iki model de belirli aralıklarda iyi uyum sağlamış, ancak bazı tahminlerde büyük sapmalar gerçekleşmiş ve yüksek değerlerde ve bazı düşük değerlerde sapmalar göstermişlerdir. Lasso ve DVR, veri seti üzerindeki tahminlerde daha geniş dağılımlar göstermiş ve ideal doğrudan daha fazla sapmalar sergilemişlerdir.

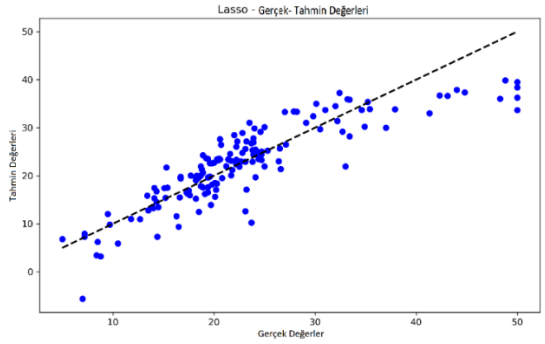
Makine öğrenimi modellerinde artık değerler ve tahmin değerleri grafiği modelin tahmin ettiği değerler ile gerçek değerler arasındaki farkları analiz ederek modelin doğruluğunu ifade etmektedir. Modelden beklenen artıkların sıfır çizgisine yakın ve rastgele bir dağılım göstermesidir. Şekil 7-11' de modellere ait artık değerler ve tahmin değerleri grafiğinde eğitim ve test verilerinin genel olarak sıfır çizgisine yakın olduğu görülmektedir. CatBoost modelinde artık değerler çoğunlukla ±2.5 aralığında kümelenmekte olması modelin daha tutarlı tahminler yaptığını göstermektedir. DVR, Karar Ağaçları ve Lasso modellerinde de eğitim ve test verileri için artık değerler sıfır çizgisine yakın yayılım göstermektedir. Ancak bazı büyük sapmalar da mevcuttur. Bu sapmalar modelin bazı durumlarda yüksek hata oranlarına sahip olduğu anlamına gelmektedir.



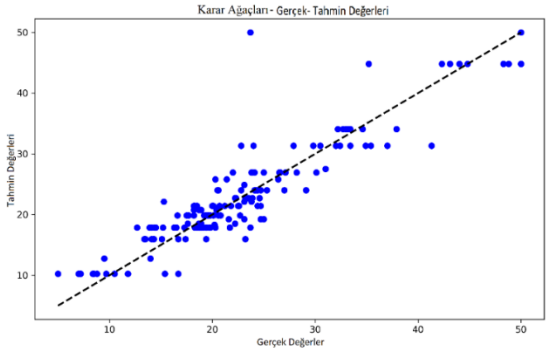
Şekil 2. CatBoost modeline ait gerçek- tahmin değerleri grafiği



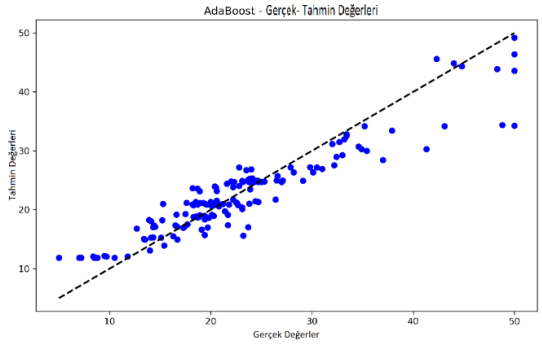
Şekil 3. DVR modeline ait gerçek- tahmin değerleri grafiği



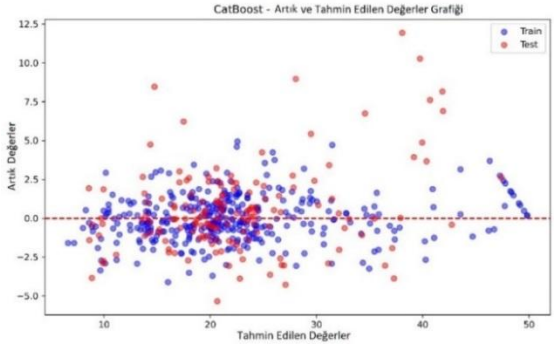
Şekil 4. Lasso modeline ait gerçek- tahmin değerleri grafiği



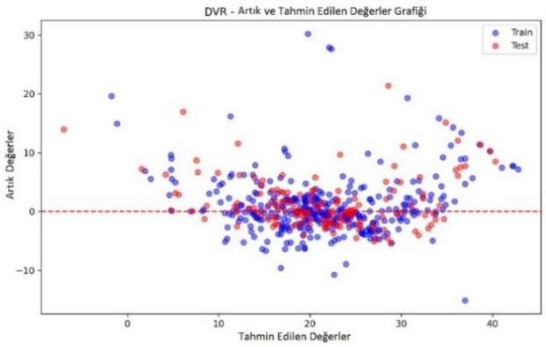
Şekil 5. Karar Ağaçları modeline ait gerçek- tahmin değerleri grafiği



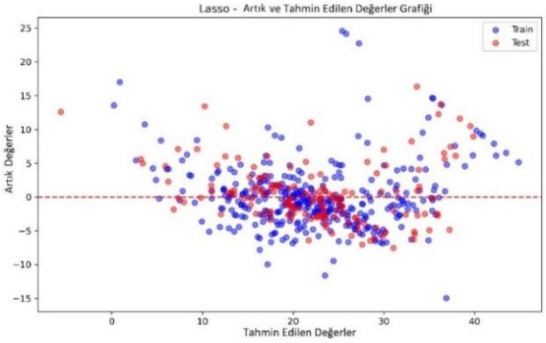
Şekil 6. AdaBoost modeline ait gerçek- tahmin değerleri grafiği



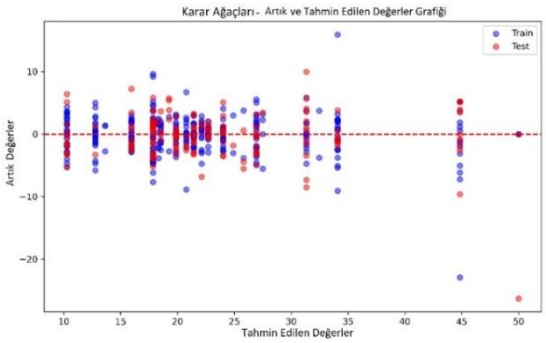
Şekil 7. CatBoost modeli için eğitim ve test verilerine ait artık değerler grafiği



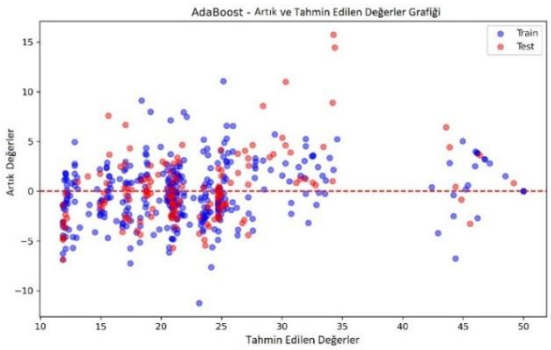
Şekil 8. DVR modeli için eğitim ve test verilerine ait artık değerler grafiği



Şekil 9. Lasso modeli için eğitim ve test verilerine ait artık değerler grafiği



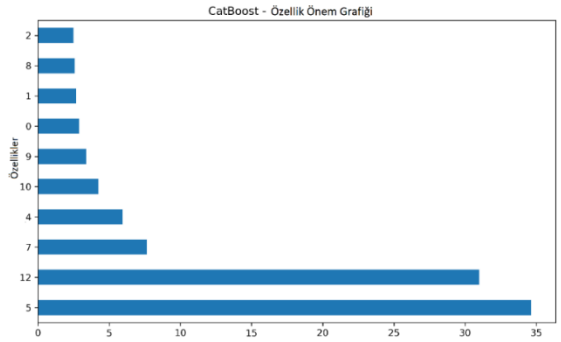
Şekil 10. Karar Ağaçları modeli için eğitim ve test verilerine ait artık değerler grafiği



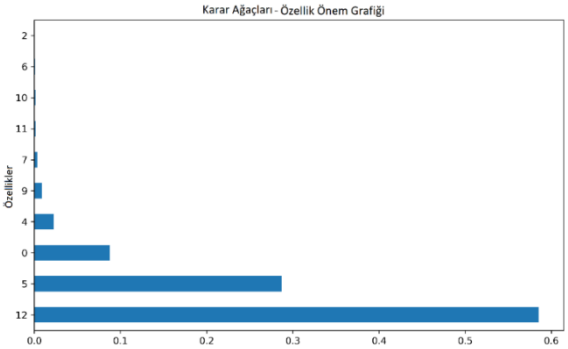
Şekil 11. AdaBoost modeli için eğitim ve test verilerine ait artık değerler grafiği

Makine öğrenme uygulamalarında önem dereceleri her bir özelliğin hedef değişkene olan etkisinin gücünü ve önemini ifade etmektedir. Şekil 12-14'te özellik önem grafiği oluşturulabilen modellere ait grafikler verilmiştir.

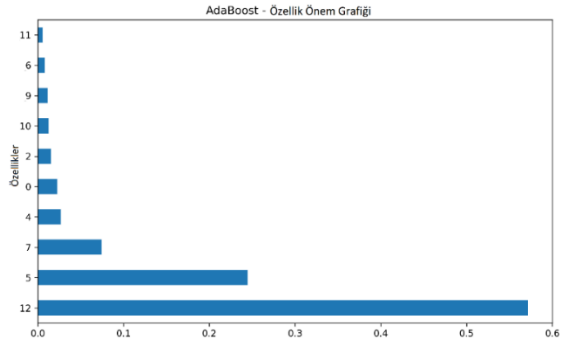
CatBoost, Karar Ağaçları ve AdaBoost'a ait özellik önem grafikleri incelendiğinde model tahminlerinde özellik-5 (Konut başına ortalama oda sayısı) ve özellik-12'nin (Düşük statüye sahip nüfus yüzdesi) en yüksek öneme sahip olduğu görülmektedir. Modeller bu özellikleri diğer özelliklere kıyasla hedef değişkeni tahmin etmede daha fazla kullanıyor anlamına gelmektedir. Modellerin karar sürecinde daha az etkili olan özellik-2 (2350 m²'den büyük arsalara sahip konut alanının oranı) ve özellik-11'in (kasabadaki siyahi nüfus oranını içeren bir ölçüt) öneminin daha düşük olduğu görülmektedir.



Şekil 12. CatBoost modeline ait seçilen özelliklerin önem grafiği



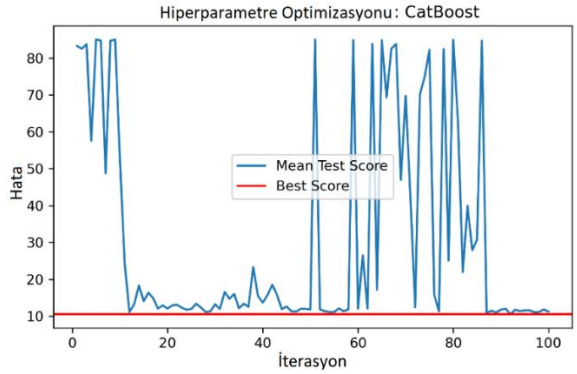
Şekil 13. Karar Ağaçları modeline ait seçilen özelliklerin önem grafiği



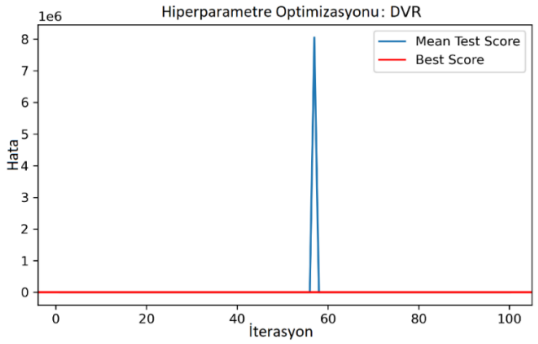
Şekil 14. AdaBoost modeline ait seçilen özelliklerin önem grafiği

Hiperparametre optimizasyonu, makine öğrenimi modellerinde performansı en üst düzeye çıkarmak amacı ile en uygun hiperparametre değerlerini belirleme sürecidir. Şekil 15-19'da verilen grafikler hiperparametre kombinasyonlarının her bir iterasyonda test edilmesiyle elde edilen hata değerlerini göstermektedir.

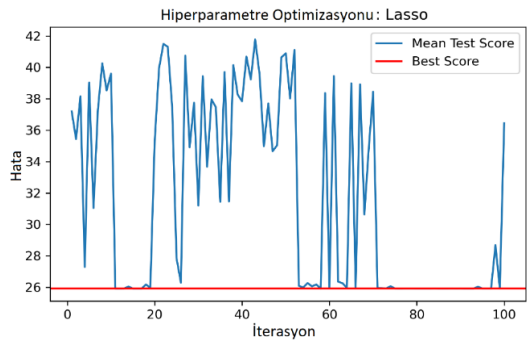
Bu optimizasyon grafiklerine göre CatBoost ve AdaBoost modellerinde dalgalanmalar olmasına rağmen diğer modellere kıyasla düşük hata seviyelerinde olduğunu ve hiperparametrelerin etkin bir şekilde optimize edildiğini göstermektedir. DVR ise tutarlılık göstermesine karşın daha yüksek hata seviyesi ile diğer modellerden daha zayıf bir performans sergilemiştir.



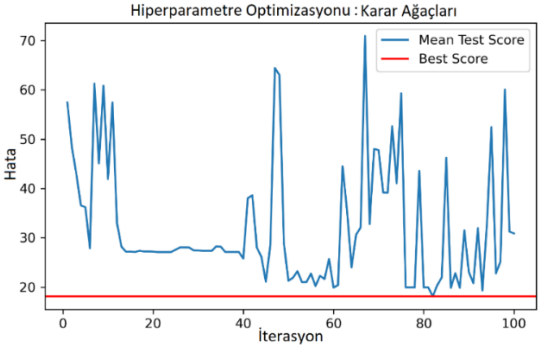
Şekil 15. CatBoost modeline ait hiperparametre optimizasyon hata değerleri grafiği



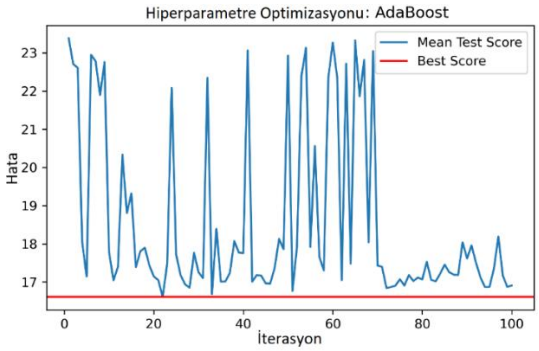
Şekil 16. DVR modeline ait hiperparametre optimizasyon hata değerleri grafiği



Şekil 17. Lasso modeline ait hiperparametre optimizasyon hata değerleri grafiği

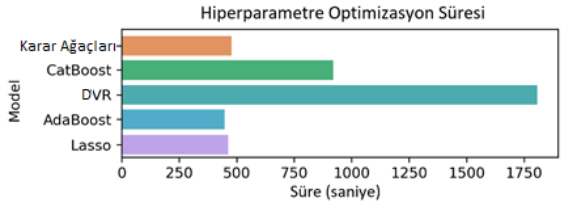


Şekil 18. Karar Ağaçları modeline ait hiperparametre optimizasyon hata değerleri grafiği



Şekil 19. AdaBoost modeline ait hiperparametre optimizasyon hata değerleri grafiği

Hiperparametre optimizasyonu süresi, modelin etkili kullanımı açısından önemli bir faktördür. Uzun süreli optimizasyonlar büyük veri setlerinde modelin kullanılabilirliğini etkilemektedir. Şekil 20' de eğitim verileri için gerçekleştirilen analizde kullanılan regresyon modellerine ait hiperparametre optimizasyon süreleri bulunmaktadır. Şekil incelendiğinde DVR modelinin en uzun optimizasyon süresine sahip olduğu görülmektedir.



Şekil 20. Kullanılan modellere ait hiperparametre optimizasyon süreleri

Eğitim ve test olarak ayrılan veriler için kullanılan her regresyon modeline ait hata metriklerinin karşılaştırılması Tablo 4'de verilmiştir. Hem eğitim hem de test veri setleri için karesel ortalama hatasında (MSE) daha düşük değer elde ederek en başarılı model CatBoost iken, en yüksek hata oranı ile başarısız olan model ise DVR modelidir. Ortalama mutlak hata (MAE), en düşük hata oranı ile CatBoost modeline ait iken Lasso ve yine DVR modeli en yüksek hata oranı ile diğer modellere kıyasla zayıf performans sergilemektedirler. R^2 değerlerinde 1'e yakınlığı ile CatBoost modelinin hem eğitim hem de test veri setindeki performansının oldukça güçlü olduğu görülmektedir. Buna karşılık DVR modeli ise 1 değerinden en uzak değeri elde etmesi sebebi ile en düşük başarıyı sergilemektedir. Ortalama mutlak yüzdesel hata (MAPE) değerlerinde CatBoost en düşük yüzde hata oranına sahiptir. Bu hata oranı tahmin edilen değerlerin gerçek değerlere yüzde bazında oldukça yakın olduğunu göstermektedir. Fiyata bağlı fark oranı (PRD) değerlerinde CatBoost modelinin hem eğitim hem de test veri setindeki değerinin 1'e yakınlığı ile test setinde en iyi doğruluğu sağlayan modeldir. En düşük COD değerlerine sahip olan CatBoost'un hem eğitim hem de test verilerinde en iyi performansı göstererek en tutarlı ve güvenilir tahminleri ürettiği görülmektedir.

Tablo 4. Kullanılan regresyon modelleri için hata metrikleri

Model	Karar Ağaçları	CatBoost	DVR	AdaBoost	Lasso
Eğitim-MSE	9.25689	2.16538	25.57186	8.74242	22.84782
Test-MSE	13.91379	8.00561	21.97877	12.04399	21.09454
Eğitim-MAE	2.11411	1.14457	3.15629	2.32317	3.29109
Test-MAE	2.51355	1.94250	3.13806	2.50424	3.38462
Eğitim-R^2	0.89027	0.97433	0.69686	0.89636	0.72915
Test-R^2	0.83468	0.90488	0.73885	0.85689	0.74936
Eğitim-MAPE	0.11248	0.06040	0.15546	0.13017	0.16631
Test-MAPE	0.12750	0.08882	0.14458	0.12772	0.15712
Eğitim-PRD	1.02448	1.01203	0.99570	1.04990	1.03650
Test-PRD	1.03786	1.00077	0.93858	1.03773	0.98178
Eğitim-COD	5.96330	4.31963	20.21973	6.78529	19.82453
Test-COD	6.87365	5.16742	19.01744	7.36941	19.10098

Hiperparametre optimizasyonu eğitim verisi üzerinde gerçekleştirilmiş olmasına rağmen, daha önce kullanılmamış test verileriyle yapılan değerlendirmelerle elde edilen hata metriklerinin de modellerin yüksek genelleme kabiliyetine ve tutarlı performansa sahip olduğunu göstermektedir. Analizde kullanılan beş regresyon modeli arasında CatBoost hem eğitim hem de test veri setlerinde düşük hata metrikleriyle en başarılı model olarak öne çıkmaktadır. MSE, MAE, R^2, MAPE, COD ve PRD değerlerinde CatBoost'un güçlü performansı, modelin yüksek genelleme kapasitesine sahip olduğunu göstermektedir.

4 Sonuçlar

Taşınmaz değerleme sürecinde kullanılan parametrelerin doğru ve tutarlı bir şekilde değerlendirilebilmesi için makine öğrenme modelleri yaygın olarak kullanılmaktadır. Bu çalışma, literatürde yaygın olarak kullanılan farklı makine öğrenme modellerinin yanı sıra modern yöntemlerden olan CatBoost algoritmasını da kullanarak taşınmaz değerleme süreçlerinde daha doğru ve etkili sonuçlar elde etmeyi hedeflemiştir. Çalışma güncel modellerden olan CatBoost

regresyonu modelini kullanarak ve hata metriklerini detaylı analiz etmesi ile literatürden farklılaşmaktadır.

Analizler sonucunda CatBoost modeli eğitim ve test veri setlerinde en düşük hata oranları ve en yüksek doğruluk doğruluk değerleri elde edilmiştir. CatBoost regresyonunda eğitim ve test verilerinde R^2 (0,97/0,90) ve PRD (1,01/1,00) değerleri 1'e en yakın ve MSE (2,16/8,00), MAE (1,14/1,94), MAPE (0,06/0,08) ve COD (4,31/5,16) değerleri için 0'a en yakın sonuçları sağlamaktadır. Bu durumda elde edilen eğitim ve test verileri sonuçlarına göre optimize edilmiş CatBoost regresyonu en başarılı sonuçlara ulaşmıştır. Destek vektör regresyonunda ise eğitim ve test verilerinde R^2 (0,69/0,73) ve PRD (0,99/0,93) değerleri 1'e ve MSE (25,57/21,97), MAE (3,15/3,13), MAPE (0,15/0,14) ve COD (20,21/19,01) değerleri 0'a en uzak sonuçları sağlamaktadır. Bu durumda DVR elde edilen eğitim ve test verisi sonuçlarına göre en düşük sonuçları elde etmiştir.

Çalışmanın bulguları hiperparametre optimizasyonunun CatBoost algoritması üzerindeki olumlu etkisini doğrulamakta ve literatürdeki diğer çalışmalardan elde edilen sonuçlarla uyum göstermektedir. Literatürle yapılan karşılaştırmalar CatBoost'un geleneksel regresyon yöntemlerine kıyasla daha başarılı olduğunu ve taşınmaz değerleme sürecinde tercih edilmesi gereken modellerden biri olduğunu göstermektedir.

Sonuç olarak CatBoost algoritmasının daha etkin çalıştığı, kategorik değişkenleri ön işlemeye gerek kalmadan doğrudan işleyebildiği, aşırı öğrenmeye karşı direnç gösterdiği, diğer birçok makine öğrenme yöntemlerine kıyasla daha hızlı olduğu hem eğitim hem test verisi üzerinde tutarlı sonuçlar üreterek genelleme gücünün yüksek olduğu ve hiperparametre optimizasyonu sayesinde model performansını artırdığı gözlemlenmiştir. Bu özellikleriyle CatBoost taşınmaz değerleme sürecinde daha güvenilir ve hassas tahminler sunarak diğer yöntemlere kıyasla üstün performans sergilemektedir. Ayrıca CatBoost'un farklı veri setleri ve coğrafi bölgelerde benzer doğruluk sağlayabileceği öngörülmektedir. Modelin sahip olduğu güçlü genelleme kabiliyeti farklı taşınmaz değerleme uygulamalarında ve farklı ölçeklerde kullanılabilirliğini desteklemektedir. Ancak, model performansının farklı veri yapıları ve bölgesel değişkenlere bağlı olarak değişebileceği göz önünde bulundurularak modelin parametre ayarlarının bölgesel ihtiyaçlara göre optimize edilmesi gerekmektedir.

Çıkar çatışması

Yazarlar çıkar çatışması olmadığını beyan etmektedir.

Benzerlik oranı (iThenticate): %8

Kaynaklar

[1] B. Demirel, A. Yelek, H. M. Alağaş ve T. Eren, Taşınmaz değerleme kriterlerinin belirlenmesi ve kriterlerin önem derecelerinin çok ölçütlü karar verme yöntemi ile hesaplanması, Kırıkkale Üniversitesi Sosyal Bilimler Dergisi, 8 (2), 665-682,2018.

[2] S. Yalpır, "Enhancement of parcel valuation with adaptive artificial neural network modeling," Artificial Intelligence Review, cilt 49, ss. 393–405, 2018.

[3] P. Çakır ve F. A. Sesli, Arsa vasıflı taşınmazların değerine etki eden faktörlerin ve bu faktörlerin önem sıralarının belirlenmesi. Harita Teknolojileri Elektronik Dergisi, 5(3), 1-16, 2013.

[4] M. Türkan, A. Bozdağ, A. E. Karkınlı ve A. G. Ulucan, Kent ölçeğinde konutlara ilişkin toplu değer değişiminin makine öğrenim algoritmaları ile analizi. Türkiye Arazi Yönetimi Dergisi, 5(2), 66-77, 2023. https://doi.org/10.51765/tayod.1275671

[5] N. Chuhan, House price prediction based on different models of machine learning, Applied and Computational Engineering, 49, 47-57, 2024. https://doi.org/10.54254/2755-2721/49/20241058

[6] A. Hazer, A. Bozdağ ve Ü. H. Atasever, Hiper-optimize edilmiş makine öğrenim teknikleri ile taşınmaz değerlemesi, Yozgat Kenti örneği, Geomatik, 9(3), 299-312, 2024. https://doi.org/10.29128/geomatik.1454915

[7] W. Yijia and Z. Qiaotong, House price prediction based on machine learning: A Case of King County, Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development, Atlantis Press, pp. 1547-1555, 2022.

[8] K. Baur, M. Rosenfelder, B. Lutz, Automated real estate valuation with machine learning models using property descriptions, Expert Systems with Applications, Vol 213, Part C, 119147, 2023. https://doi.org/10.1016/j.eswa.2022.119147

[9] M. Hernes, P. Tutak, M. Nadolny, A. Mazurek, Real estate valuation using machine learning, Procedia Computer Science, Vol 246, Pages 4592-4599, 2024. https://doi.org/10.1016/j.procs.2024.09.323

[10] P. Jafary, D. Shojaei, A. Rajabifard, T. Ngo, Automated land valuation models: A comparative study of four machine learning and deep learning methods based on a comprehensive range of influential factors, Cities, Vol 151, 105115, 2024. https://doi.org/10.1016/j.cities.2024.105115

[11] K. Konhauser, T. Werner, Uncovering the financial impact of energy-efficient building characteristics with eXplainable artificial intelligence, Applied Energy, Vol 374, 123960, 2024. https://doi.org/10.1016/j.apenergy.2024.123960

[12] U. Grzybowska, H. Dudek, A. Wojewódzka-Wiewiórska, Socioeconomic factors associated with household overcrowding in the Visegrad Group countries – analysis based on machine learning approach, Procedia Computer Science, Vol 246, Pages 4441-4450, 2024. https://doi.org/10.1016/j.procs.2024.09.294

[13] Kalliola, J., Kapočiūtė-Dzikienė, J., and Damaševičius, R., Neural network hyperparameter optimization for prediction of real estate prices in Helsinki, PeerJ computer science, 7, 444, 2021.

[14] A. Ç. Aydınoğlu ve R. Bovkır, İ. Çölkesen, Toplu taşınmaz değerlemede makine öğrenme algoritmalarının kullanımı ve konumsal/konumsal olmayan özniteliklerin tahmin doğruluğuna etkilerinin

karşılaştırılması, Journal of Geodesy and Geoinformation, Vol. 10, 63-83, 2023.
[15] Boston şehrine ait veri seti. https://www.kaggle.com/datasets/schirmerchad/bostonhoustingmlnd
[16] A. Coşkuner, Gayrimenkul Yatırım Ortaklıklarında Kârlılık Belirleyicilerinin Veri Madenciliği Yöntemleri İle Tahminlemesi, Ankara, 2024.
[17] A. Dorogush, V. Ershov, A. Gulin, CatBoost: gradient boosting with categorical features support, 2018. https://doi.org/10.48550/arXiv.1810.11363
[18] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush and A. Gulin, CatBoost: unbiased boosting with categorical features, Proceedings of the 32nd International Conference on Neural Information Processing Systems, Curran Associates Inc., pp. 6639–6649, Montréal, Canada, 2018.
[19] H. Nasiri, A. Tohry and H. R. Heidari, Modeling industrial hydrocyclone operational variables by SHAP-CatBoost - A "conscious lab" approach, Powder Technology, 420, 2023. https://doi.org/10.1016/j.powtec.2023.118416
[20] C. Cortes, V. Vapnik, Support-Vector Networks, Machine Learning, 20, 273-297, 1995. https://doi.org/10.1007/BF00994018
[21] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Statistics and Computing, 14, 199–222,(2004). https://doi.org/10.1023/B:STCO.0000035301.49549.88
[22] G. James, D. Witten, T. Hastie and R. Tibshirani, An Introduction to Statistical Learning, 2013.
[23] M. Apaydın, M. Yumuş, A. Değirmenci, ve Ö. Karal, Evaluation of air temperature with machine learning regression methods using Seoul City meteorological data, Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 28 (5),737–747, 2022.
[24] S. Rosset, H. Zou, T. Hastie, Multi-class AdaBoost, Statistics and its interface, 2(3), 2006. https://doi.org/10.4310/SII.2009.v2.n3.a8
[25] T. T. Joy, S. Rana, S. Gupta and S. Venkatesh, Hyperparameter tuning for big data using Bayesian optimization, 23rd International Conference on Pattern Recognition (ICPR), 2574-2579, 2016.
[26] Feurer, M., Springenberg, J., ve Hutter, F., Initializing bayesian hyperparameter optimization via metalearning, In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 29, No. 1, 2015.
[27] M. Khosravi, S. B. Arif, A. Ghaseminejad, H. Tohidi, H. Shabanian, Performance evaluation of machine learning regressors for estimating, Real Estate House Prices. Preprints 2022. https://doi.org/10.20944/preprints202209.0341.v1
[28] A. Anghel, N. Papandreou, T. P. A. de Palma and H. Pozidis, Benchmarking and optimization of gradient boosting decision tree algorithms, Workshop on Systems for ML and Open Source Software at NeurIPS, Canada, 2018. https://doi.org/10.48550/arXiv.1809.04559
[29] L. Yang and A. Shami, On hyperparameter optimization of machine learning algorithms: theory and practice, Neurocomputing, 2022. https://doi.org/10.48550/arXiv.2007.15745
[30] K. Šehić, A. Gramfort, J. Salmon and L. Nardi, LassoBench: A high-dimensional hyperparameter optimization benchmark suite for lasso, AutoML Conference,2022. https://doi.org/10.48550/arXiv.2111.02790
[31] A. Duran ve H. Bakır, Hiperparametreleri ayarlanmış makine öğrenimi algoritmalarını kullanarak android sistemlerde kötü amaçlı yazılım tespiti, USBTU 2(1): 1-19, 2023.
[32] R. Gao and Z. Liu, An Improved adaboost algorithm for hyperparameter optimization, Journal of Physics: Conference Series, Vol. 1631, 2020.