NÖHÜ Müh. Bilim. Derg. / NOHU J. Eng. Sci., 2022; 11(4), 838-845

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

Pırlanta fiyat tahmini için regresyon modellerinin karşılaştırmalı analizi

Comparative analysis of regression models for predicting diamond price

Merve Asil[1], Gülfem Işıklar Alptekin[2,*]

[1] *Galatasaray University, Graduate School of Science and Engineering, 34349, İstanbul Turkey*
[2] *Galatasaray University, Computer Engineering Department, 34349, İstanbul, Turkey*

Öz

Bilişim dünyasındaki gelişmeler ile artan veri hacmi ve çeşitliliği ile birlikte, hayatımıza büyük veri kavramı girmiş ve beraberinde birçok zorluğu da peşinde getirmiştir. Verinin işlenebilirliği büyük bir önem kazanmış ve güncel kullanılan bazı veri işleme yöntemlerinin performansı yetersiz gelmeye başlamıştır. Büyük veri analizinde yapay zekâ ve makine öğrenmesi teknikleri kullanılarak bu sorunlar çözülmeye çalışılmakta ve gün geçtikçe daha etkin çözümler bulan algoritmalar önerilmeye devam edilmektedir. Bu çalışmanın amacı, iyi bilinen ve sıklıkla kullanılan regresyon algoritmalarını bir veri kümesi üzerinde çalıştırmak, performans sonuçlarını karşılaştırarak en iyi sonuç verenleri sunmaktır. Makalede pırlantaların kesimi, rengi, berraklığı ve fiyatı gibi özellikleri barındıran açık bir veri kümesi kullanılmıştır. Verilerin ön işlemesi yapılmış, tanımlayıcı analiz gerçekleştirilmiş ve fiyatlarının tahmini için farklı regresyon modelleri hem ilkel hem de optimize edilmiş halleriyle çalıştırılmıştır. Regresyon modelleri içinden diğerlerine kıyasla daha düşük RMSE ve daha yüksek r^2 değerleri GBM modelleri (özellikle Light GBM) ve rassal orman algoritmasında alınmıştır.

Anahtar kelimeler: Regresyon modelleri, Makine öğrenmesi, Fiyat tahmini, Doğrusal regresyon, Doğrusal olmayan regresyon.

Abstract

With the developments in informatics and the increasing volume and diversity of data, the concept of big data has entered our lives and brought many challenges with it. The usefulness of data has gained great importance and the performance of some commonly used data processing methods have begun to be insufficient. These problems have been tried to be solved by using artificial intelligence and machine learning techniques that find more effective solutions. The aim of this study is to run well-known and frequently used machine learning algorithms on a public dataset, to compare their performance results comparatively, and present the best performant ones. A public data on diamonds is preprocessed, descriptive analysis is performed, and various regression models to predict the corresponding prices are run, both in their primitive and optimized forms, GBM models (especially Light GBM) and random forest algorithm have the lowest RMSE values and highest r^2 values compared to other models.

Keywords: Regression models, Machine learning, Price prediction, Linear regression, Non-linear regression.

1 Giriş

Bir regresyon modeli, bir veya daha fazla bağımsız değişken ile bir bağımlı/hedef değişken arasındaki ilişkiyi tanımlayan bir fonksiyon ortaya koyar. Regresyon fonksiyonu birçok tahmin tipinin esasını oluşturur ve hedef değişken üzerindeki etkilerin görülmesini sağlar. Bu sayede, sektör göstergelerine dayalı olarak önemli iş kararları alınabilir. Verilen iş kararları ve sonuçları arasındaki ilişkiler uzun vadeli olarak belirlenebilirse, ileride alınan kararların çok daha yerinde olması sağlanabilir. Eğer fonksiyondaki girdiler ve çıktılar arasındaki ilişki düz bir çizgi ile ifade edilebiliyorsa, doğrusal bir regresyon olduğu söylenir. Gerçek dünyada gözlemlenmesi en basit olan regresyon tipi budur. Eğer girdiler ve çıktılar arasında doğrusal bir ilişki kurulamıyorsa, doğrusal olmayan bir regresyon modeli olduğu varsayılır. Doğrusal olmayan fonksiyonlar, çok çeşitli eğrilere uyabilir. Bu yüksek sayıda eğri adayı arasından, veriler için en uygun formun seçilmesi gerekir. Bu makalede, bu seçim süreci ele alınmış ve veri kümesi olarak Kaggle'da yer alan 'Diamonds' veri kümesi seçilmiştir [1]. Fiyat tahmini, birçok farklı ürün ve servis için en sık gereksinim duyulan iş problemlerinin başında yer alır. Dolayısıyla, pırlanta gibi nispeten çok özelliği barındıran ve bu özelliklere göre fiyatı değişen bir ürünün fiyatını tahmin edecek regresyon fonksiyonu üzerinde çalışmanın, regresyon modellerinin karşılaştırması için uygun olacağı düşünülmüştür.

Akademik yazında, bu makalede gerçekleştirilen çalışmaya en yakın özellikteki çalışmalar Tablo 1'de özetlenmiştir. İlk grup olarak, Kaggle'da açık olarak bulunan pırlantalarla ilgili özellikler ve fiyatları içeren 'Diamonds' veri kümesini kullanan çalışmalar verilmiştir. Bunlardan biri, birbiriyle kıyaslamak için sekiz farklı gözetimli öğrenme algoritması kullanmıştır [2].

* Sorumlu yazar / Corresponding author, e-posta / e-mail: gisiklar@gsu.edu.tr (G. I. Alptekin)
Geliş / Recieved: 16.03.2022 Kabul Kabul / Accepted: 18.07.2022 Yayımlanma / Published: 14.10.2022
doi: 10.28948/ngmuh.1088916

Tablo 1. Akademik yazındaki benzer çalışmalar

Çalışma	Kullanılan veri kümesi	Performans ölçütleri	En performanslı algoritma
[2]	Diamonds (Kaggle)	Doğruluk, RMSE	Rassal orman
[3]	Diamonds (Kaggle)	RMSE	Rassal orman ve kolektif modeller
[4]	Diamonds (Kaggle)	Doğruluk	Önerilen kolektif model
[5]	Diamonds (Kaggle)	r^2, MAE, RMSE	CatBoost, rassal orman, XGB
[6]	2. el araç fiyatları (Kaggle)	MSE	Rassal orman
[7]	Konut fiyatları	RMSE, MSE	GBM
[8]	Türkiye 2. el araç fiyatları	RMSE, MAE	Doğrusal regresyon
[9]	Sinema bilet fiyatları (bookmyshow.com)	RMSE	Rassal orman
[10]	ABD ev fiyatları (Kaggle)	RMSE	CatBoost
Kendi çalışmamız	Diamonds (Kaggle)	r^2, RMSE	Light GBM, rassal orman

Tablo 2. Veri kümesinden örnek olarak verilen beş kayıt

	Carat	Cut	Color	Clarity	Depth	Table	Price	x	y	z
0	0.23	İdeal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75

Bu çalışmada, modeller optimize edilmemiş ve doğruluk değerleri ve RMSE hata oranına göre yaptıkları sıralamada en yüksek performanslı algoritma olarak rassal orman algoritmasını seçmişlerdir. Aynı veri kümesini kullanan diğer bir çalışmada [3], algoritmalar veriler temizlenmeden uygulanmıştır ve rassal orman veya kolektif modellerinin kullanımı önerilmiştir. Aynı veri kümesi üzerinde çalışılırken elde edilen bu sonuç farkları, veri kümesinin temizlenmesinin yarattığı fark olarak değerlendirilebilir. Diğer bir çalışma [4], daha az sayıda algoritmayı kendi önerdikleri kolektif bir modelle karşılaştırmışlardır. Kısa bir ön çalışma şeklinde yazılan bu makalede, kendi önerdikleri modelin diğerlerinden daha iyi doğruluk oranı verdiğini söylemişlerdir. Pırlanta veri kümesi üzerinde yapılan en kapsamlı çalışmada [5], uygulama biçimleri verilmese de, en yüksek doğruluk oranı ve düşük RMSE değeri veren CatBoost regresyon modeli olmuştur. Onu takip eden algoritmalar, bizim çalışmamızda da olduğu gibi rassal orman ve XGBoost regresyondur. Algoritmaları r^2 (r-kare), RMSE (*Root Mean Squared Error*) ve MAE (*Mean Absolute Error*) değerlerine göre değerlendirmişlerdir. Aynı veri kümesi üzerinde yapılan çalışmaların sonuçları arasındaki bu farklar, modelin optimize edilmesinin ve verinin temizlenmesinin, yapılan hata oranını düşüreceğini ve en performanslı algoritma seçimini etkileyeceğini göstermiştir. Fiyat tahmini yaparken regresyon modelleri, çok çeşitli sektörlerde kullanılmaktadır. İkinci el araçların fiyatlarını belirlemek için gerçekleştirilen bir çalışmada [6], performans ölçütü olarak MSE hata değeri alınmış ve en az hatayı veren rassal orman algoritması olmuştur. Diğer bir çalışmada [7], tüketicilerin finansal beklentileri uyarınca konut fiyatlarının tahmininde regresyon modelleri kullanılmıştır. Farklı algoritmaları RMSE ve MSE (*Mean Squared Error*) ölçütlerine göre karşılaştırmışlar ve GBM algoritmasının en yüksek doğruluk ve en az hata oranı verdiğini saptamışlardır. Türkiye'deki ikinci el araç fiyatlarını tahmin etmek için yapılan diğer bir çalışmada, Türkiye içinde toplanılan bir veri kümesinden faydalanılmış e doğrusal regresyon modeli kullanılmıştır [8]. Modelin performansı RMSE ve MAE hata değerleri ile değerlendirilmiş ve modelin tatmin edici sonuçlar verdiği söylenmiştir. Sinema biletlerinin fiyatlarını dinamik olarak belirleyebilmek amacıyla, veri kümeleme ve regresyon yöntemlerinin bir arada kullanıldığı bir yaklaşım önerilmiştir [9]. Önerilen hibrit modelin performansını ölçmek için RMSE değeri kullanılmış ve rassal orman modelinin en az hatalı sonuçları verdiğini gösterilmiştir. Konutların birçok farklı özelliğini kullanarak fiyatlarını düzgün belirleme problemi için gerçekleştirilen bir diğer çalışmada, CatBoost regresyon algoritması en düşük RMSE değerini vermiştir [10]. Bu çalışmada, ABD'deki ev fiyatlarını içeren açık bir veri kümesi kullanılmıştır.

Akademik yazındaki çalışmalarla kıyaslandığında, bu çalışmanın katkıları şu şekilde özetlenebilir:

- Seçilen veri kümesi üzerinde modelleri çalıştırmadan önce, veri ön işleme adımları, aykırı gözlem analizi ve tanımlayıcı analiz uygulanmıştır.
- Veri kümesinin ön işlenmesinden algoritma sonuçlarının elde edilmesine kadar olan süreç ayrıntılı şekilde sunulduğundan, bu konuda çalışacak araştırmacılar için uygulanabilir bir örnek olmuştur.
- Kullanılan modeller hem ilkel hem de optimize edilmiş şekilde uygulanmış ve sonuçlar karşılaştırılmıştır.
- Aynı veri kümesiyle çalışırken kullanılan verinin temizlenmesi ve kullanılan modelin optimize edilmesinin alınan performans ölçütlerini olumlu yönde değiştirebileceği gösterilmiştir.

Tablo 1'de anlatılan benzer çalışmaların hiçbirinde veri ön işleme adımları, aykırı gözlem analizi, tanımlayıcı analiz gibi adımlar sunulmamıştır. Algoritmalar doğrudan uygulanmış ve seçilen performans ölçütlerine göre en iyi sonuç verenler saptanmıştır. Kendi çalışmamızdaki sonuçlara bakıldığında (Tablo 4), akademik yazındakilerin bir kısmına benzer şekilde, en düşük RMSE ve en yüksek r^2 değerlerinden biri rassal orman ve Light GBM için bulunmuştur.

Aynı veri kümesi üzerinde çalışılırken elde edilen bu sonuç farkları, veri kümesinin temizlenmesinin ve modelin optimize edilmesinin yarattığı fark olarak değerlendirilebilir. Makalenin 2. bölümünde veri kümesi ve özelliklerinden bahsedildikten sonra, tüm kullanılan yöntemler ve alınan sayısal sonuçlar özetlenmiştir. 3. bölümde karşılaştırma sonuçları verilmiş, yorumlanmış ve makale sonuçlandırılmıştır.

2 Veri kümesi ve yöntem

Bu bölümde kullanılan veri kümesi, içindeki bağımlı ve bağımsız değişkenler, veri ön işleme adımında kullanılan yöntemler ve sonuçları, tanımlayıcı analiz çalışmaları ve karşılaştırılan regresyon modellerinden en iyi sonuç veren iki yöntem ve sonuçları sunulmuştur. Kullanılan veri kümesi açık bir veri kümesi olduğu için [1], yapılan çalışmaların tekrar edilebilmesi ve öğrenme/ karşılaştırma için kullanılabilmesi amacıyla, adımlar ayrıntılı şekilde verilmiştir.

2.1 Veri kümesi ve değişkenleri

Mücevherat sektöründe kullanılan bir pahalı taş olan pırlantanın fiyatı, renginden, berraklığına kadar birçok özelliğine bağlı olarak değişmektedir. Bu çalışmada, internet üzerinden ulaşılabilecek [1] bir pırlanta veri kümesi kullanılmıştır. Veri kümesi toplam 10 değişkenden oluşmaktadır. Bunlar arasında, pırlanta fiyatlarını gösteren '*price*' sürekli nicel bir değişken ve '*cut*', '*color*', '*clarity*' gibi nitel, '*carat*', '*depth*', '*table*', '*x*', '*y*' ve 'z' gibi nicel değişkenler bulunmaktadır. '*Carat*' pırlantanın karat ağırlığı, '*cut*' pırlantanın kesim kalitesi (Fair-Good-Very Good-Premium-Ideal değerlerini alır.), '*color*' pırlantanın rengi (En kötüden en iyiye doğru J, I, H, G, F, E, D değerlerini alır.), '*clarity*' pırlantanın berraklığı (En kötüden en iyiye doğru I1, SI2, SI1, VS2, VS1, VVS2, VVS1, IF değerlerini alır.), '*depth*' pırlantanın yüksekliğinin ortalama kuşak çapına bölünmesi, '*table*' pırlantanın genişliğinin ortalama çapının yüzdesi, '*x*' pırlantanın uzunluğu (mm), '*y*' pırlantanın genişliği (mm) ve 'z' pırlantanın yüksekliği (mm) şeklinde tanımlanmıştır. Veri kümesinden örnek olarak ilk beş kayıt, Tablo 2'de verilmiştir. Veri kümesi 3 MB büyüklüğünde olup, toplamda 53940 kayıt içermektedir.

2.2 Veri ön işleme adımları

Çalışmaya başlarken, verinin ön işleme adımının bir parçası olan, nicel değişkenlere ait ortalama ve medyan değerlerine bakılarak, aykırı gözlem analizi yapılmıştır (Tablo 3). Tablo 3'te en fazla farklılık '*price*' değişkeninde saptanmıştır. Şekil 1'de, bu değişkene ait keman grafiği çizilmiş ve fiyatlar yükseldikçe, aykırılıkların arttığı görülmüştür.

Fiyat değişkeni için aykırı gözlemleri düzenlemek için kullanılabilecek üç farklı yöntem uygulanmış ve her üç durum için de ilkel (parametreler üzerinde işlem yapılmamış) RMSE ve r^2 değerleri hesaplanarak, en iyi sonuç verenler seçilmiştir. r^2 değeri 0-1 arasında belirlenmekte olup, iki değişken arasındaki korelasyonu açıklar. r^2 değeri 1'e ne kadar yakınsa, seçilen regresyon eğrisi veri ile o derece uyumludur. RMSE, bir model veya bir tahminci tarafından tahmin edilen değerler ile gözlemlenen değerleri karşılaştırır ve iki veri kümesi arasındaki farka bakarak ne kadar hata olduğunu ölçer.

Tablo 3. Nicel değişkenlere ait ortalama ve medyan değerleri

Değişken	Ortalama	Medyan
carat	0.80	0.70
depth	61.75	61.80
table	57.46	57.00
price	3932.80	2401.00
x	5.73	5.70
y	5.73	5.71
z	3.54	3.53

Hesaplanan RMSE değeri ne kadar küçükse, gözlemlenen ve tahmin edilen değerler birbirine o kadar yakın demektir. Aykırı gözlemleri düzenlemek için kullanılabilecek yöntemler aşağıda özetlenmiştir:

i. Aykırıları silmek: Bağımlı değişken olan '*price*' için yapılan incelemede, 53940 gözlem arasından 3540 tanesinin aykırı olduğu tespit edilmiştir. Bunlar silinerek, veri kümesi 50400 gözleme düşürülmüştür.

ii. Medyan ile doldurmak: Bu işlemin yapılma amacı, aykırı '*price*' değerlerini kendi medyan değeri ile doldurarak veri kaybı yaşamamaktır. Bağımlı değişkenin medyan değeri 2401 olarak hesaplanmıştır ve bütün aykırı değişken değerleri yerine bu değer konmuştur.

iii. Baskılama yöntemi: Bu yöntemde, '*price*' değişkeninin aykırı değerleri için tespit edilen alt (Denklem (1)) ve üst sınırlara (Denklem (2)) göre işlem yapılmaktadır. Bağımlı değişken finansal bir değer olduğundan, keman grafiğinde (Şekil 1) de görüldüğü gibi aykırı gözlemler üst sınırda bulunmaktadır. Inter Quartile Range (IQR) hesaplamak için,

veri kümesinde birinci çeyreklik (Q1) ve üçüncü çeyreklik (Q3) hesaplanarak, aradaki fark bulunmuştur.

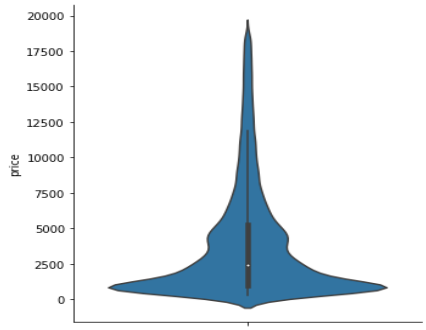
$$alt\_sinir = Q1 - 1.5 * IQR \quad (1)$$

$$ust\_sinir = Q3 + 1.5 * IQR \quad (2)$$

Baskılama yönteminde, aykırı '*price*' değişkenlerini doldurmak için hesaplanan üst sınır değeri olan 8192.5 kullanılmıştır.

2.3 *Veri kümesi üzerinde tanımlayıcı analizi*

'Diamonds' isimli veri kümesi eksiği olmayan ve nispeten temiz veri içerdiği için, araştırmamızın ilk aşaması veri içindeki kalıpları, ilişkileri ve anormallikleri bulmayı içeren tanımlayıcı veri analizine ayrılmıştır. Tablo 3'te listelenen değişkenlerin veri kümesi içindeki frekansları incelenmiştir. Örneğin, veri kümesindeki kayıtlara '*cut*' değişkenine göre bakıldığında, en fazla 'ideal' kesimde pırlanta olduğu, '*color*' değişkenine göre bakıldığından ise en fazla 'G' renginde pırlanta olduğu görülmüştür. Her bir değişkene ait frekanslar birer histogram şeklinde çizilmiş ve ardından, bu değişkenlerin ikili ve üçlü kombinasyonları uyarınca fiyat değişikliklerine bakılmıştır. '*cut*' ve '*color*' değişkenleri uyarınca '*price*' değişkeninin değişimini gösteren örnek bir grafik, Şekil 2'de verilmiştir.

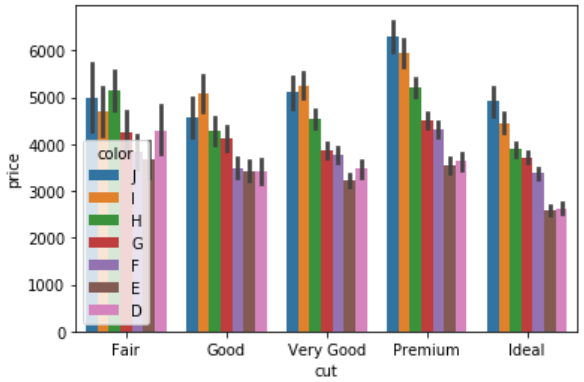


Şekil 1. '*price*' değişkenine ait keman grafiği

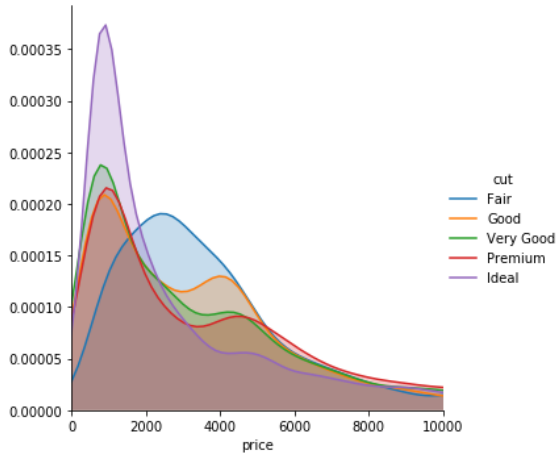
Bu grafiklerden şuna benzer sonuçlar çıkarmak mümkündür: Hemen hemen tüm kesimler ('*cut*') için, '*J*' rengine sahip bir pırlanta en yüksek fiyata sahiptir; ardından '*I*' rengi gelmektedir. Yine bu incelemelere ek olarak, değişkenler için olasılık yoğunluk fonksiyonu grafikleri de çizilmiştir (Şekil 3). Şekil 3'te de, 'ideal' kesime sahip olan pırlantaların ağırlığı gözlemlenebilmektedir.

Bu değişimlere veya frekanslara bakmanın sebebi, algoritmaları çalıştırmadan önce eldeki veriyi daha iyi tanımaktır. Tüm bu grafikleri çizmek için Python programlama dili kullanılmıştır. Tanımlayıcı analizde işe yarayan araçlardan bir diğeri de çift grafiklerdir. Çift grafikler hem tek değişkenin dağılımını hem de iki nicel değişken arasındaki ilişkileri görmek için kullanılır. Çift grafik oluşturulurken histogram ve serpme grafiklerinden yararlanılır. Şekil 4'te, değişkenlerin bazıları arasında (örneğin '*price*' ve '*carat*') yüksek korelasyon olduğu görülmektedir. Gerçekten de bir pırlantanın karat ağırlığının artması ile pırlanta fiyatında da belirli bir miktar artış olması beklenir. Diğer bir yüksek korelasyon, '*price*' değişkeni ile '*x*', '*y*', '*z*' değişkenleri arasında görülen pozitif yönlü korelasyondur.

Şekil 4'teki çift grafiği, en fazla ilişkili olduğu sağlanan değişkenlerle çizilen Şekil 5'teki ısı haritasıyla birleştirildiğinde, değişkenler arasındaki korelasyonlar daha ayrıntılı şekilde saptanmıştır. Şekil 5'e göre, örneğin pırlantanın fiyatı ile karat değeri arasında 0.92 değerinde pozitif yönlü yüksek bir ilişki olduğu gözlemlenmiştir. Yine benzer şekilde, karat değerinin *'x', 'y', 'z'* değerlerini pozitif yönlü 0.95 ve daha yüksek seviyede etkilediği görülmektedir. *'x', 'y', 'z'* değişkenlerinin hem kendi arasındaki korelasyonlar yüksektir; hem de bağımlı değişken olan '*price*' ile de pozitif yönlü yüksek bir korelasyon saptanmıştır. Bu ilk incelemeler ışığında, fiyata etkisi yüksek olan değişkenler ayrı ayrı incelenmiştir. Örnek olarak, Şekil 6'da '*carat*' ve '*price*' değişkenlerinin bir arada değerlendirildiği ve boyut olarak '*clarity*' değişkeninin eklendiği saçılım (*scatter plot*) grafiği verilmiştir. Şekil 6'da 'I1' berraklığına sahip pırlantaların karat değerinin, fiyata karşı daha duyarlı olduğu dikkat çekmektedir. Yer kısıtı sebebiyle hepsi makalede verilemese de çizilen farklı bir saçılım grafiğinde, 'Fair' kesimine sahip 'I1' berraklığında pırlantanın da, fiyata karşı oldukça duyarlı olduğu gözlemlenmiştir.



Şekil 2. '*cut*' ve '*color*' değişkenleri uyarınca '*price*' değişkeninin değişimi



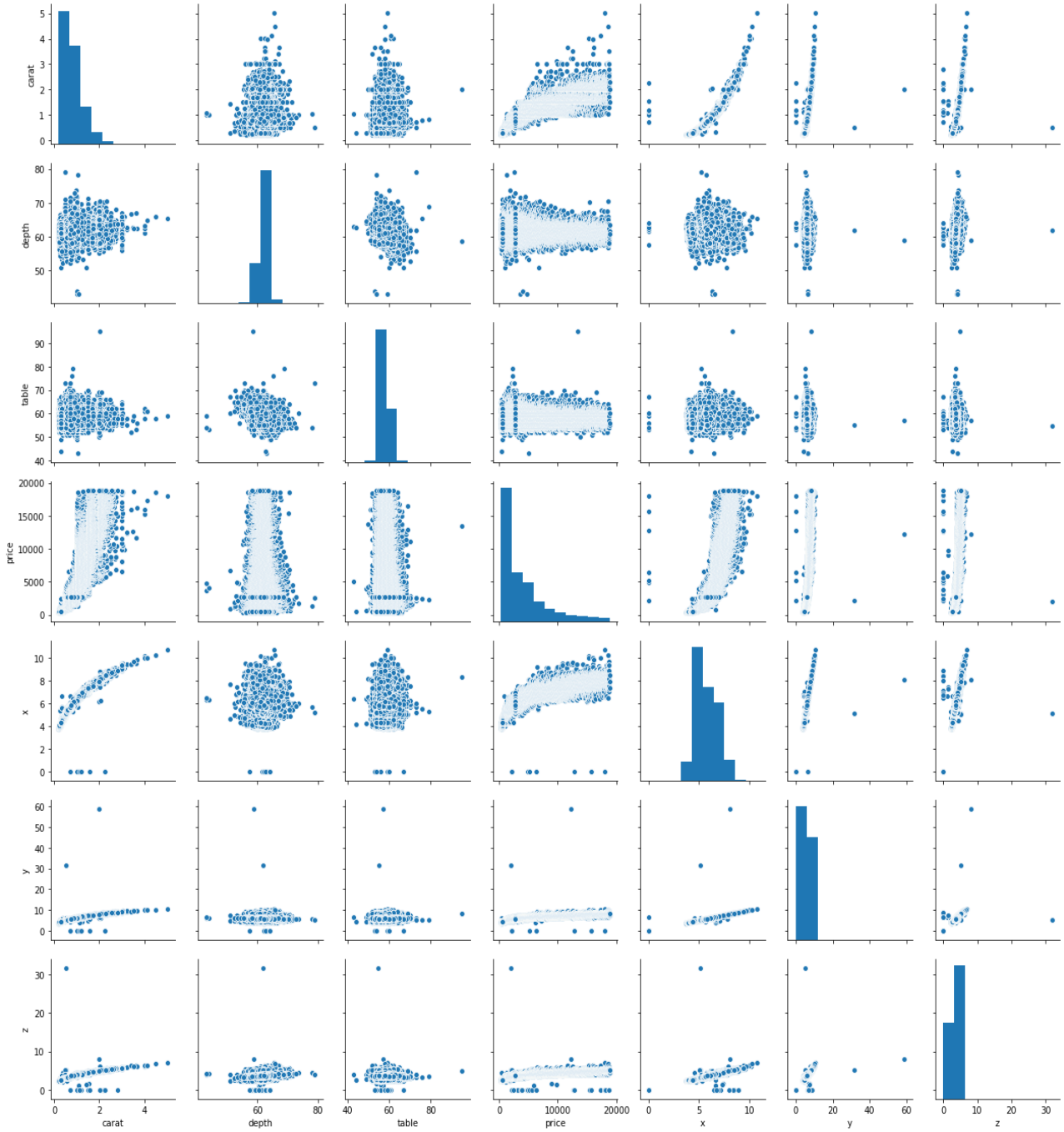
Şekil 3. '*cut*' ve '*price*' değişkenlerinin bir arada değerlendirildiği olasılık yoğunluk fonksiyonu grafiği

2.4 *Veri kümesi üzerinde regresyon modelleri*

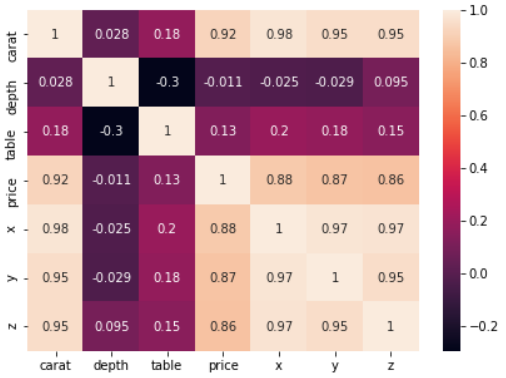
Regresyon modelleri, iki ya da daha fazla nicel değişken arasındaki ilişkiyi ölçmek için kullanılan analiz yöntemidir. Bir regresyon modeli kurulduğunda, modelin veriye ne kadar 'iyi' uyduğunu bilmek gereklidir. Diğer bir deyişle, modelin bağımlı değişkenin değerini tahmin etmek için, bağımsız değişkenleri ne derece etkin kullanabildiğini anlamak gereklidir. İstatistikte bunu ölçmek için sıklıkla kullanılan iki ölçüt RMSE ve r^2 değerleridir.

RMSE, bir model tarafından tahmin edilen değerler ile gözlemlenen değerleri karşılaştırır ve iki veri kümesi arasındaki farka bakarak ne kadar hata olduğunu ölçer. r^2, bağımsız değişkenin bağımlı değişkeni etkilediği durumların yüzdesel hacmini vererek, aralarındaki ilişkinin gücünü 0-1 aralığında gösterir. Dolayısıyla, RMSE'in mümkün olduğunca küçük, r^2 değerinin ise mümkün olduğunca büyük olması beklenmektedir. Bu çalışmada, 'Diamonds' veri kümesi için en uygun regresyon modelini belirlerken, RMSE ve r^2 ölçütlerinin ikisini bir arada değerlendirmek tercih edilmiştir [11].

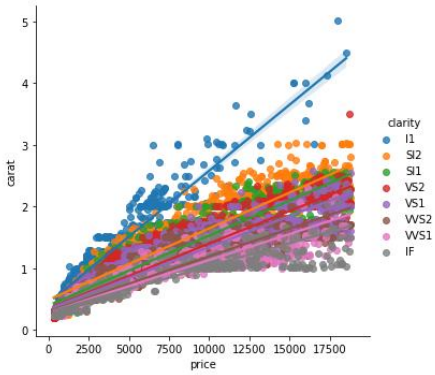
Eldeki veriye en uygun regresyon modelini belirlemek için, doğrusal ve doğrusal olmayan 12 regresyon modelleri denenmiştir.



Şekil 4. Değişkenlerin birbirleri arasındaki ilişkiyi gösteren çift grafiği



Şekil 5. Değişkenlerin birbirleri arasındaki ilişkiyi gösteren ısı grafiği



Şekil 6. '*carat*' ve '*price*' değişkenlerinin bir arada değerlendirildiği ve boyut olarak '*clarity*' değişkeninin eklendiği saçılım grafiği

Veri ön işleme bölümünde elde edilen üç farklı veri kümesine (*df_sil*, *df_median* ve *df_baski*) ek olarak, herhangi bir veri ön işleme sürecinden geçirilmeden, sadece üzerinde kategorik değişkenlerin kukla değişkenlere dönüştürülme işlemi uygulanan ham veri kümesi (*df_ilk*) kullanılarak, ilkel RMSE ve r^2 değerleri hesaplanmıştır. Bunu yaparken, RMSE değeri hesaplanmadan önce her bir veri kümesi *hold-out* yöntemi uygulanarak, %75 eğitim, %25 test olmak üzere iki parçaya ayrılmıştır. Modeller eğitim veri kümesinde parametrelerinde değişim yapılmaksızın (ilkel) kurulmuş ve test veri kümesi ile kurulan modellere ait RMSE ve r^2 değerleri hesaplanmıştır. Elde edilen dört veri kümesi kullanılarak ayrı ayrı tüm regresyon modellerine ait RMSE değerleri hesaplanmış, bu veri kümelerinden en düşük RMSE değerinin *df_sil* (aykırı değerlerin veri kümesinden silindiği)'e ait olduğu görülmüştür. Bu işlemle, aykırı gözlemlerin makine öğrenmesi algoritmaları kullanılarak yapılan tahminlerinde yanlılığa ve tahminler üzerinde sapmalara neden olduğu, anomali yaratan değerlerin veri kümesinden silindiğinde daha düşük hata değerleri elde edilmesiyle bir kez daha doğrulanmıştır. Bir sonraki adımda, hiperparametreler üzerinde optimizasyon işlemleri bu veri kümesi (*df_sil*) üzerinde gerçekleştirilmiştir. Aykırı değerlerin silindiği veri kümesinin tüm modeller için RMSE ve r^2 değerleri Tablo 4'te gösterilmiştir.

Modelleri optimize etme işleminde k-katlı çapraz doğrulama (*k-fold cross validation*) yönteminden yararlanılmıştır. Algoritmaların hiperparametrelerine rassal bazı değerler verilip ('*k*' kez), farklı parametre değerleri ile çaprazlama işlemi yapılarak test edilmiştir. İşlemin sonunda en iyi (yani en düşük) RMSE değerini veren parametre seçilip, model tekrar kurulmuştur. Bu yeni kurulan modele de ayarlanmış model denilmiştir. Kullanılan tüm algoritmalar için, veri kümesinde bulunan gözlem sayıları da göz önünde bulundurularak kat (*k*) 10 olarak seçilmiştir. Her bir regresyon algoritması için optimum hiperparametre değerini bulmakta, *scikit-learn* kütüphanesinde yer alan *GridSearchCV* fonksiyonundan yararlanılmıştır. 'Diamonds' veri kümesi üzerinde farklı regresyon modelleri denenmiş ve en başarılı iki tanesi RMSE ve r^2 değerleri ile Tablo 4'te gösterilmiştir. Light GBM modeli, diğer *boosting* modelleri ile karşılaştırıldığında, işlem hızının ve tahmin oranının yüksek olması ve bununla beraber RAM kullanımının düşük olması ile ayrışmaktadır. Aralarında en iyi performans, Light GBM modelinden elde edilmiştir. Light GBM'den sonra performansı yüksek olan algoritma rassal orman olmuştur. Bu sonuç, Tablo 1'de verilen benzer çalışmalar ile de uyumlu çıkmıştır. Bu iki modele ait uygulama ayrıntıları aşağıda verilmiştir

2.4.1 Rassal ormanlar (random forest - RF)

Kurulan ilkel *rf_model*'in hazır hiperparametre değerleri aşağıda verilmiştir:

RandomForestRegressor(bootstrap = True, ccp_alpha = 0.0, criterion = 'mse', max_depth = None, max_features = 'auto', max_leaf_nodes = None, max_samples = None, min_impurity_decrease = 0.0, min_impurity_split = None, min_samples_leaf = 1, min_samples_split = 2, min_weight_fraction_leaf = 0.0, n_estimators = 100, n_jobs = None, oob_score = False, random_state = 42, verbose = 0, warm_start = False)

Kurulan ilkel model üzerinden hesaplanan ilkel hata değerini optimum hale getirmek için çapraz doğrulama kullanıldığında, optimum değerler olarak, *max_depth* 500, *max_features* 20, *min_samples_split* 40, *n_estimators* 1000 olarak bulunmuştur. Bu değerler kullanılıp, eğitim veri kümesiyle rassal ormanlar algoritmasıyla model tekrar kurulduğunda (*rf_tuned_model*), test veri kümesi için modelin hesapladığı RMSE değeri 374.797 olmuştur. Bu modelin optimizasyonunun hata oranını düşürmediği, ama az da olsa r^2 değerinde bir artış yarattığı görülmüştür. Karar ağaçlarında olduğu gibi, bağımsız değişken önem seviyeleri rassal ormanlar modelinde de mevcuttur. Önem seviyelerine bakıldığında, yine '*y*' değişkeninin üstünlüğü göze çarpmıştır.

2.4.2 Light GBM

XGB algoritmasının eğitim ve tahmin performansını arttırabilmesi için 2017 yılında Microsoft tarafından LightGBM algoritması geliştirilmiştir [12]. Kurulan ilkel modelin (*lgb_model*) hazır hiperparametre değerleri aşağıda verilmiştir:

LGBMRegressor(boosting_type = 'gbdt', class_weight = None, colsample_bytree = 1.0, importance_type = 'split', learning_rate = 0.1, max_depth = -1, min_child_samples = 20, min_child_weight = 0.001, min_split_gain = 0.0, n_estimators = 100, n_jobs=-1, num_leaves = 31, objective = None, random_state = None, reg_alpha = 0.0, reg_lambda = 0.0, silent = True, subsample = 1.0, subsample_for_bin = 200000, subsample_freq = 0)

Tüm ağaç modellerinde olduğu gibi LightGBM modeli için de model öğrenme süresine etki eden bağımsız değişkenlerin sayısı, bağımlı değişkeni etkileme oranına değerleri üzerinden yapılmıştır. Kurulan modelin optimizasyonu için GBM ve XGB algoritmalarındaki parametrelerle benzerlik gösteren hiperparametreler seçilmiştir. Bu doğrultuda, *colsample_bytree* ve *subsample* gibi örnek alt ağaç kümeleri için kullanılan parametreler hazır değerleriyle kullanılmış olup, üzerinde çapraz doğrulama yapılıp optimumu bulunan hiperparametreler ve değerleri şöyle olmuştur: *learning rate* 0.1, *max_depth* 500 ve *n_estimators* 500. Bu optimum değerlerle, LightGBM algoritmasıyla model tekrar kurulduğunda oluşan *lgb_tuned_model*in test veri kümesi için hesapladığı RMSE değeri, %5.7 oranında bir azalma göstererek 352.187 olmuştur (Tablo 4).

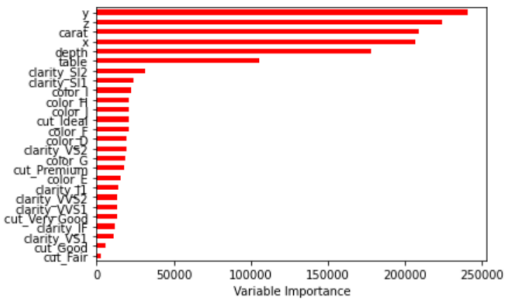
Algoritmaların performanslarını daha yakından incelemek amacıyla, Light GBM ile rassal orman algoritması, her bir *k* değerinde verdiği RMSE ve r^2 değeri ile birlikte Tablo 5'te sunulmuştur.

3 Tartışma ve sonuç

Kullanılan doğrusal modeller için, optimizasyon yapılmamıştır. Tüm optimize edilmiş algoritmalar arasında en düşük hata oranı Light GBM'de görülmüştür. Light GBM'in aynı zamanda en yüksek r^2 değerlerinden birini verdiği de gözlemlenmiştir. Light GBM için önem sırasındaki ilk beş değişken şu şekilde çıkmıştır: *y*, *z*, *carat*, *x*, *depth* (Şekil 7). Akademik yazında, 'Diamonds' veri kümesi üzerinde çeşitli algoritmalar kullanmış çalışmalar, Tablo 1'de verilmiştir. Bu çalışmaların hiçbirinde yöntemlerin uygulama ayrıntıları verilmemiştir; dolayısıyla bu çalışmaların tekrar edilmesi mümkün değildir.

Bu makaledeki çalışma, verinin ön işlemeden geçirilmesi, tanımlayıcı analizin yapılması, modellerin hem ilkel, hem optimize edilmiş halleriyle hata oranlarının ölçülmesi ve uygulama ayrıntıları verilerek tekrar edilebilirliği sağlaması bakımlarından diğerlerinden ayrılmaktadır. Aynı veri kümesi üzerinde yapılan çalışmalarda elde edilen sonuç farklılıkları ise, modelin optimize edilmesinin ve verinin temizlenmesinin, elde edilen hata oranını düşüreceğini ve en performanslı algoritmanın seçimini doğrudan değiştirebileceğini ispatlamaktadır.

İleriki çalışmalarda, kullanılan veri çoğaltılarak, genişletilmiş veri kümesinin sonuçlara etkisine bakılabilir. Veri çoğaltmak için öğrenme veri kümesinde yer alan gözlemlerin şansa bağlı olarak seçilerek minör ve majör sınıfsal denge kurulana kadar çoğaltılması yöntemi seçilebilir. Bu şekilde veri kümesinin yetersizliği problemi bertaraf edilebilir. Bunun dışında, kullanılan özelliklerin sayısını azaltmak ve en etkin özellik alt kümesini bulabilmek için, Temel Bileşenler Analizi veya Ortak Faktör Analizi yöntemlerinden faydalanılabilir.



Şekil 7. *lgb_model*indeki bağımsız değişkenlerin önem sırası

Tablo 4. En başarılı regresyon modellerinin ilkel ve optimize edilmiş RMSE ve r^2 değerleri ile karşılaştırılması

Regresyon modeli	İlkel modelin RMSE değeri	İlkel modelin r^2 değeri	Optimize edilmiş modelin RMSE değeri	Optimize edilmiş değerin r^2 değeri	RMSE değişimi (%)	r^2 değişimi (%)
Rassal Orman	370.480	0.981	374.797	0.980	1.2	-0.1
Light GBM	373.616	0.983	352.187	0.984	-5.7	0.1

Tablo 5. Seçilen iki algoritmaya ait k-kat değeri değişimine göre RMSE ve r^2 değerleri

	k=1		k=2		k=3		k=4		k=5		k=6	
	RMSE	r^2	RMSE	r^2	RMSE	r^2	RMSE	r^2	RMSE	r^2	RMSE	r^2
Rassal Orman	363.271	0.982	374.189	0.981	366.297	0.982	407.796	0.978	392.53	0.979	364.670	0.982
Light GBM	367.612	0.981	376.857	0.980	366.862	0.982	403.239	0.978	388.805	0.980	365.469	0.982
	k=7		k=8		k=9		k=10		Ortalama			
	RMSE	r^2	RMSE	r^2	RMSE	r^2	RMSE	r^2	RMSE	r^2		
Rassal Orman	377.991	0.981	368.635	0.982	373.723	0.981	358.873	0.983	374.797	0.981		
Light GBM	371.041	0.982	370.386	0.982	364.896	0.982	360.992	0.983	373.616	0.981		

Çıkar çatışması

Yazarlar çıkar çatışması olmadığını beyan etmektedir.

Benzerlik oranı (iThenticate): %6

Kaynaklar

[1] Kaggle Diamonds Dataset. https://www.kaggle.com/shivam2503/diamonds, Accessed 02 March 2022.

[2] G. Sharma, V. Tripathi, M. Mahajan and A. K. Srivastava, Comparative analysis of supervised models for diamond price prediction. Proceedings of 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 1019-1022, Uttar Prades, India, 2021.

[3] W. Alsuraihi, E. Al-hazmi, K. Bawazeer and H. Alghamdi, Machine learning algorithms for diamond price prediction. Proceedings of 2nd ACM International Conference on Image, Video and Signal Processing (IVSP '20), pp. 150–154, Singapour, 2020.

[4] A. C. Pandey, S. Misra and M. Saxena, Gold and diamond price prediction using enhanced ensemble learning. Proceedings of 12th International Conference on Contemporary Computing (IC3), pp. 1-4, Noida, India, 2019.

[5] H. Mihir, M. I. Patel, S. Jani and R. Gajjar, Diamond price prediction using machine learning. Proceedings of 2nd IEEE International Conference on Communication, Computing and Industry 4.0 (C2I4), pp. 1-5, Bangalore, India, 2021.

[6] N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya and P. Boonpou, Prediction of prices for used car by using regression models. Proceedings of 5th International Conference on Business and Industrial Research (ICBIR), pp. 115-119, Thailand, 2018.

[7] C. R. Madhuri, G. Anuradha and M. V. Pujitha, House price prediction using regression techniques: A comparative study. Proceedings of International Conference on Smart Structures and Systems (ICSSS), pp. 1-5, Madras, India, 2019.

[8] M. C. Satıoğlu, Y. Ar ve B. Tuğrul, Automobile price prediction in Turkey marketplace with linear regression. Proceedings of 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 329-333, Turkey, 2021.

[9] V. Gupta, K. Singh, S. K. Arjaria and B. Biswas, Dynamic pricing in movie tickets using regression techniques. Proceedings of International Conference on Advanced Computation and Telecommunication (ICACAT), pp. 1-4, India, 2018.

[10] G. K. Kumar, D. M. Rani, N. Koppula and S. Ashraf, Prediction of house price using machine learning algorithms. Proceedings of 5th International Conference on Trends in Electronics and Informatics (ICOEI), pp. 1268-1271, India, 2021.

[11] G. Yin, F. J. I. Alazzawi, S. Mironov, F. Reegu, A. S. El-Shafay, M. L. Rahman, C. H. Su, Y. Z. Lu and H. C. Nguyen, Machine learning method for simulation of adsorption separation: Comparisons of model's performance in predicting equilibrium concentrations, Arabian Journal of Chemistry. 15 (3), 103612, 1-10, 2022. https://doi.org/10.1016/j.arabjc.2021.103612.

[12] LightGBM, https://www.microsoft.com/en-us/research/project/lightgbm/, Accessed 03 March 2022.