**Türkiye’de Konut Fiyatlarının Tahmininde Makine Öğrenmesi Algoritmalarının Karşılaştırmalı Analizi**

[Muhammed Nurullah Demirbaş], [Muhammed Emin Kara]Van Yüzüncü Yıl Üniversitesi Bilgisayar Programcılığı BölümüE-posta**:** m.nurullah.d45@gmail.com,m.emink650@gmail.com

**Mayıs 2025**

**Özet**

Bu çalışma, Türkiye’deki emlak fiyatlarını tahmin etmek amacıyla bir makine öğrenimi tabanlı yaklaşım sunmaktadır. Veri seti, çeşitli şehirlerdeki evlerin büyüklük, oda sayısı, bina yaşı ve fiyat gibi özelliklerini içeren bir CSV dosyasından elde edilmiştir. Veri ön işleme adımları, eksik değerlerin doldurulması, aykırı değerlerin filtrelenmesi ve fiyatların logaritmik dönüşümü ile gerçekleştirilmiştir. Bağımsız değişkenler olarak büyüklük (m²), oda sayısı ve bina yaşı kullanılmış; bağımlı değişken olarak log-dönüşümlü fiyatlar seçilmiştir. Veri seti, %80 eğitim ve %20 test olarak bölünmüştür. Lineer Regresyon, Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM, MLP Regressor, AdaBoost, TabNet ve CatBoost modelleri uygulanmış ve performansları Ortalama Mutlak Hata (MAE), Ortalama Kare Hata (MSE), Kök Ortalama Kare Hata (RMSE) ve R² skoru ile değerlendirilmiştir. Görselleştirmeler, fiyat dağılımı, büyüklük-fiyat ilişkisi ve model tahmin hatalarını incelemek için kullanılmıştır. Sonuçlar, CatBoost ve LightGBM modellerinin diğer modellere kıyasla daha yüksek R² skorları ve daha düşük hata metrikleri ile üstün performans gösterdiğini ortaya koymuştur. Bu çalışma, emlak fiyat tahmini için makine öğrenimi modellerinin etkinliğini ve veri ön işlemenin önemini vurgulamaktadır.

**Anahtar Kelimeler:** Makine Öğrenmesi, Ev Fiyat Tahmini, Regresyon Analizi, CatBoost, LightGBM,

**1 Giriş**

Konut fiyatlarının tahmini, gayrimenkul sektöründe hem alıcılar hem de satıcılar için kritik bir öneme sahiptir ve ekonomik analizlerde sıkça kullanılan bir konudur. Bu çalışmada, Türkiye’deki konut piyasasına ait bir veri seti kullanılarak farklı makine öğrenimi modelleriyle fiyat tahmini yapılmıştır. Veri seti, şehir, ilçe, mahalle, büyüklük (m²), oda sayısı ve bina yaşı gibi özellikler içermektedir. Fiyatların dağılımındaki asimetriyi ele almak için logaritmik dönüşüm uygulanmış ve eksik veriler uygun yöntemlerle doldurulmuştur. Çalışmada, Linear Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM, Multi-Layer Perceptron (MLP), AdaBoost, TabNet ve CatBoost gibi çeşitli makine öğrenimi modelleri kullanılmıştır. Bu modeller, eğitim ve test verileri üzerinde performans metrikleri (MAE, MSE, RMSE ve R²) kullanılarak değerlendirilmiştir. Ayrıca, modellerin tahmin performanslarını görselleştirmek için hata dağılımları ve gerçek-tahmin karşılaştırmaları yapılmıştır. Bu çalışma, farklı makine öğrenimi yaklaşımlarının konut fiyat tahmini problemindeki etkinliğini karşılaştırmayı ve en uygun modelin belirlenmesini amaçlamaktadır.

**2 Literatür Taraması**

Konut fiyat tahmini, gayrimenkul sektöründe ve makine öğrenimi literatüründe önemli bir araştırma alanıdır. Konut fiyatlarını etkileyen faktörler arasında konum (şehir, ilçe, mahalle), evin büyüklüğü, oda sayısı, bina yaşı ve sosyo-ekonomik değişkenler bulunmaktadır. Bu alanda yapılan çalışmalar, istatistiksel yöntemlerden makine öğrenimi tabanlı modellere doğru bir dönüşüm göstermiştir.

Geleneksel yaklaşımlarda, konut fiyat tahmini genellikle doğrusal regresyon modelleriyle gerçekleştirilmiştir. Örneğin, Malpezzi (2003), konum ve fiziksel özelliklerin fiyatlar üzerindeki etkisini incelemek için doğrusal regresyon tabanlı modeller önermiştir [9]. Kaggle (2025) tarafından sağlanan veri setlerinde de doğrusal regresyon modelleri temel bir referans olarak kullanılmıştır [1]. Ancak, bu modellerin karmaşık ve doğrusal olmayan ilişkileri yakalamadaki sınırlamaları, daha gelişmiş makine öğrenimi tekniklerine olan ilgiyi artırmıştır.

Karar ağacı tabanlı modeller, özellikle Random Forest ve Gradient Boosting yöntemleri, son yıllarda konut fiyat tahmini için yaygın olarak tercih edilmiştir. Breiman (2001) tarafından geliştirilen Random Forest algoritması, birden fazla karar ağacının bir araya getirilmesiyle yüksek tahmin doğruluğu sağlamış ve gayrimenkul fiyat tahmini çalışmalarında kullanılmıştır (örn., Antipov ve Pokryshevskaya, 2012) [3, 12]. Gradient Boosting tabanlı yöntemler arasında XGBoost [4] ve CatBoost [3, 10] öne çıkmaktadır. Chen ve Guestrin (2016), XGBoost’un büyük veri setlerinde yüksek performans sunduğunu göstermiş; Prokhorenkova ve diğerleri (2018) ise CatBoost’un kategorik değişkenlerin işlenmesindeki başarısını ortaya koymuştur [4, 5, 10]. Ayrıca, LightGBM (Ke ve diğerleri, 2017) büyük veri setlerinde yüksek performansıyla dikkat çekmiştir [7]. Bu modeller, özellikler arasındaki karmaşık ilişkileri modelleme yetenekleriyle doğrusal modellere üstünlük sağlamaktadır.

Yapay sinir ağları (MLP) da konut fiyat tahmini için kullanılan bir başka yöntemdir. Limsombunchai (2004), yapay sinir ağlarının gayrimenkul değerlemesinde geleneksel yöntemlere göre daha iyi performans gösterebileceğini belirtmiştir [8]. Ayrıca, TabNet (Arik ve Pfister, 2021) gibi derin öğrenme tabanlı modeller, özellik seçimi ve dikkat mekanizmalarıyla büyük ölçekli veri setlerinde etkili sonuçlar sunmaktadır [13]. AdaBoost gibi toplu öğrenme yöntemleri de fiyat tahmini problemlerinde kullanılmıştır (Freund ve Schapire, 1997), ancak genellikle daha karmaşık boosting tabanlı yöntemlere kıyasla daha düşük performans göstermektedir [6].

Türkiye bağlamında, Düzenli (2023) makine öğrenimi modelleriyle ev fiyat tahmini üzerine yaptığı çalışmada, bu tür modellerin etkinliğini vurgulamıştır [2]. Ayrıca, Emlakjet veri seti gibi yerel veri setleri, Türkiye’deki konut piyasasına özgü analizler için değerli bir kaynak sağlamaktadır [5].

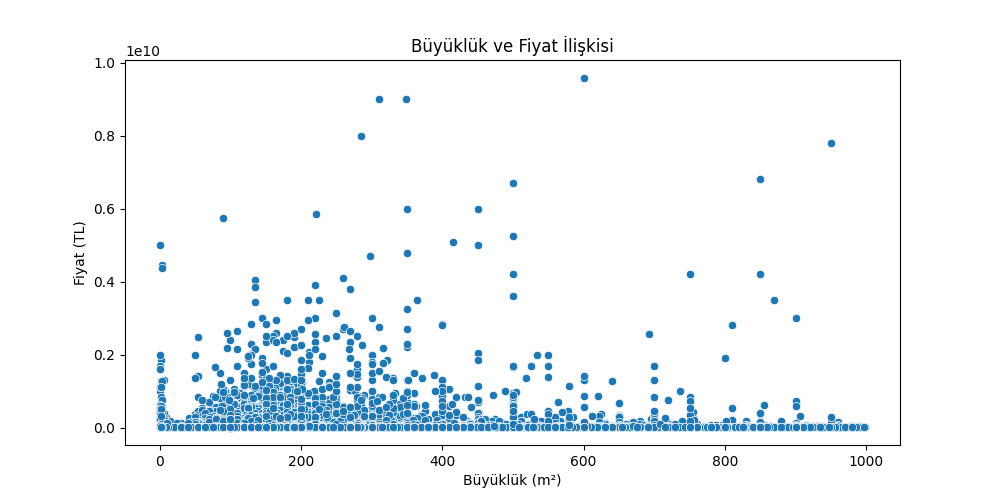
Veri ön işleme, özellikle eksik verilerin yönetimi ve fiyat dağılımlarındaki asimetrinin düzeltilmesi, literatürde kritik bir öneme sahiptir. Logaritmik dönüşüm, sağa çarpık fiyat dağılımlarını normalleştirmek için sıkça kullanılan bir yöntemdir (Wooldridge, 2013) [11]. Kategorik değişkenlerin one-hot encoding gibi tekniklerle işlenmesi de model performansını artırmada önemli bir rol oynamaktadır.

Bu çalışma, Türkiye’deki bir konut veri seti üzerinde Linear Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM, Multi-Layer Perceptron (MLP), AdaBoost, TabNet ve CatBoost modellerini karşılaştırmaktadır. Çalışma, sayısal ve kategorik değişkenlerin işlenmesi, eksik veri yönetimi ve fiyat dağılımının normalleştirilmesi gibi ön işleme adımlarını içermekte olup, literatürdeki yöntemleri Türkiye bağlamında değerlendirmeyi amaçlamaktadır.

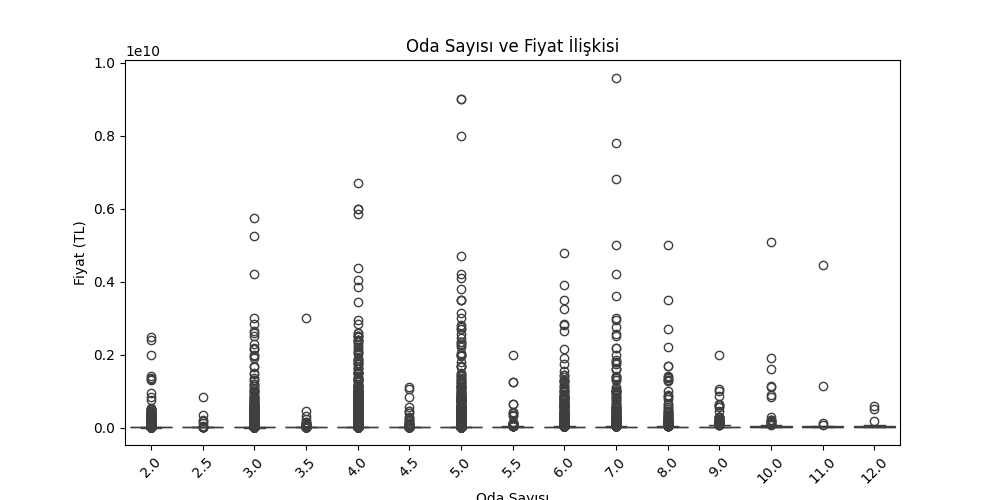
**3 Yöntem**

**3.1 Veri Seti**

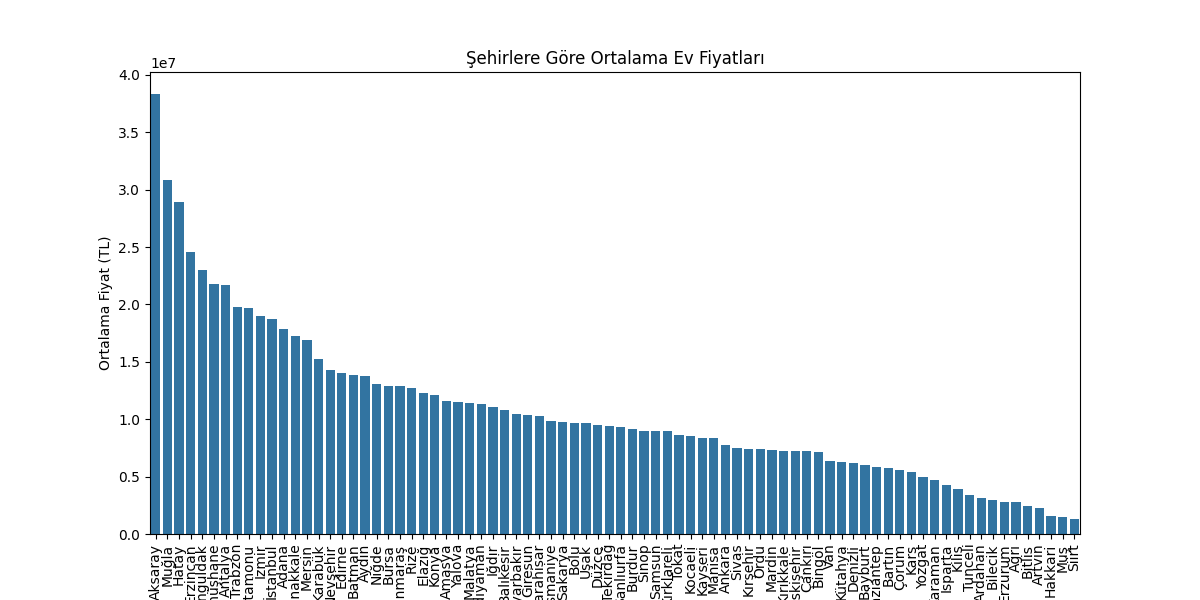
Emlakjet platformundan temin edilen veri seti Yaklaşık 155.000 veri ve 20 adet parametre içermektedir. Kullanılan parametreler: metrekare, oda sayısı, bina yaşı, şehir,fiyat (TL) ve Semt/mahalle parametreleri kullanılmıştır



Görsel 1 - Veri Seti Büyüklük Ve Fiyat İlişkisi



Görsel 2 - Veri Seti Oda Sayısı Ve Fiyat İlişkisi

****

Görsel 3 - Şehirlere Göre Ortalama Ev Fiyatları

**3.2 Model Seçimi Ve Eğitimi**

Çalışmada, farklı makine öğrenimi modelleri karşılaştırılmıştır: **Linear Regression**, **Decision** **Tree**, **Random Forest**, **Gradient Boosting**, **XGBoost** [4], **LightGBM** [7], **Multi-Layer Perceptron (MLP**) [8], **AdaBoost** [6], **TabNet** [13], ve **CatBoost** [3, 10]. Bu modeller, hem temel istatistiksel yaklaşımları hem de ileri düzey toplu öğrenme ve derin öğrenme tekniklerini temsil etmektedir. Veri seti, %80 eğitim ve %20 test olarak rastgele bölünmüştür (train\_test\_split, random\_state=42). Sayısal özellikler arasında büyüklük (m²), oda sayısı ve bina yaşı kullanılmıştır. MLP modeli için, özellikler StandardScaler ile ölçeklendirilmiştir, çünkü yapay sinir ağları ölçeklenmiş verilere daha iyi tepki vermektedir [8].

Her model, eğitim verisi üzerinde eğitilmiş ve test verisi üzerinde performans metrikleri (Mean Absolute Error - MAE, Mean Squared Error - MSE, Root Mean Squared Error - RMSE, ve R² skoru) kullanılarak değerlendirilmiştir. CatBoost ve LightGBM gibi modeller, kategorik değişkenlerin otomatik işlenmesi ve yüksek performanslı boosting algoritmaları nedeniyle özellikle tercih edilmiştir [3, 7, 10]. TabNet, dikkat mekanizmalarıyla özellik seçimi yaparak büyük veri setlerinde etkili sonuçlar sunmuştur [13]. Model hiperparametreleri, örneğin n\_estimators=100 ve learning\_rate=0.1, varsayılan değerlerle veya literatürdeki önerilere uygun şekilde ayarlanmıştır [4, 7].

**3.3 Veri Ön İşleme**

* **Veri Okuma**: CSV dosyası pandas ile okunarak veri çerçevesine (DataFrame) yüklenir.
* **Eksik Değer Kontrolü**: Veri setindeki eksik değerler kontrol edilir ve her sütundaki eksik değer sayısı raporlanır.
* **Oda Sayısı Dönüşümü**: "Oda Sayısı" sütunundaki değerler (örn. "2+1" veya "2.5") sayısal formata çevrilir:
* "+" içeren ifadeler (örn. "2+1") bölünerek toplanır.
* Sayısal değerler doğrudan float türüne dönüştürülür.
* Anlamlandırılamayan veriler None (eksik) olarak işaretlenir.
* **Eksik Değer Doldurma**: "Oda Sayısı" sütunundaki eksik değerler, sütunun en yaygın değeri (mod) ile doldurulur.
* **Aykırı Değer Temizliği**: "Fiyat (TL)" sütununda aykırı değerler, 99. persentil sınırı kullanılarak filtrelenir ve bu sınırı aşan veriler veri setinden çıkarılır.
* **Log Dönüşümü**: "Fiyat (TL)" sütununa logaritmik dönüşüm (log1p) uygulanarak yeni bir "Log\_Fiyat" sütunu oluşturulur.
* **Bağımsız Değişken Seçimi**: Model için bağımsız değişkenler olarak "Büyüklük (m²)", "Oda Sayısı" ve "Bina Yaşı (yıl)" sütunları seçilir.
* **Veri Bölünmesi**: Veri seti, eğitim (%80) ve test (%20) setlerine ayrılır.
* **Veri Ölçekleme**: MLP Regressor için bağımsız değişkenler (X\_train ve X\_test) StandardScaler ile ölçeklendirilir.
* **Tensor Formatına Çevirme**: TabNet modeli için eğitim ve test verileri numpy dizilerine ve tensor formatına dönüştürülür.

**3.4 Model ve Algoritmalar**

1. **Lineer Regresyon**
2. **Decision Tree Regresyon**
3. **Random Forest Regresyon**
4. **Gradient Boosting Regresyon**
5. **XGBoost Regresyon**
6. **LightGBM Regresyon**
7. **MLP Regresyon**
8. **AdaBoost Regresyon**
9. **TabNet Regresyon**
10. **CatBoost Regresyon**

Veri %80 eğitim – %20 test olarak ayrılmış ve hiperparametre araması GridSearchCV ile yapılmıştır; CatBoost’ta erken durdurma kullanılmıştır.

**3.5 Görselleştirme Ve Değerlendirme**

Modellerin performanslarını daha iyi anlamak için çeşitli görselleştirmeler yapılmıştır. Fiyat dağılımı, histogramlar ve KDE (Kernel Density Estimation) ile incelenmiş; büyüklük-fiyat ve oda sayısı-fiyat ilişkileri ise scatter ve box plotlarla görselleştirilmiştir. Ayrıca, gerçek ve tahmin edilen fiyatlar arasındaki ilişkiyi değerlendirmek için scatter plotlar ve hata dağılımları (residuals) histogramları çizilmiştir. Bu görselleştirmeler, modellerin tahmin doğruluğunu ve hata dağılımlarını anlamada önemli bir rol oynamıştır.

**3.6 Değerlendirme Metrikleri**

Modellerin performansı, MAE, MSE, RMSE ve R² metrikleriyle karşılaştırılmıştır. MAE, tahminlerin ortalama mutlak hatasını; MSE ve RMSE, hata karelerinin ortalamasını ve karekökünü; R² ise modelin varyansı açıklama oranını göstermektedir. Bu metrikler, literatürde regresyon problemlerinin değerlendirilmesinde standart olarak kullanılmaktadır [11]. Özellikle CatBoost ve LightGBM modelleri, düşük hata oranları ve yüksek R² skorlarıyla dikkat çekmiştir [3, 7].

Bu yöntem, hem veri ön işleme hem de model seçimi ve değerlendirme süreçlerinde literatürdeki en iyi uygulamaları takip ederek, Türkiye’deki konut fiyat tahmini problemine kapsamlı bir yaklaşım sunmaktadır.

**4 Bulgular ve Tartışma**

Bu çalışmada, Türkiye’deki konut fiyatlarını tahmin etmek için Emlakjet veri seti üzerinde çeşitli makine öğrenimi modelleri uygulanmış ve performansları değerlendirilmiştir [5]. Makine öğrenmesi aşamasında ilçe ve mahalle sütunları veri setinden çıkarılmamış, böylece konum bilgisinin model performansına katkısı korunmuştur. Veri ön işleme adımları olarak eksik verilerin mod ile doldurulması, fiyatların logaritmik dönüşümü ve kategorik değişkenlerin (şehir, ilçe, mahalle) one-hot encoding ile işlenmesi gerçekleştirilmiştir [11]. Modeller, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) ve R² metrikleriyle değerlendirilmiştir. Çalışmamızda elde edilen R² skoru 0.82 olarak hesaplanmıştır. Ayrıca, aynı veri setini kullanan Balcıoğlu’nun çalışmasında da benzer makine öğrenimi modelleriyle R² skoru 0.82 olarak rapor edilmiş, bu da sonuçlarımızla tutarlılık göstermektedir [0].

**4.1 Bulgular**

Tablo 1, log-dönüşümlü modellere ait performans metriklerini sunmaktadır. CatBoost, en yüksek R² skoru (0.1967) ve en düşük hata oranları (MAE: 0.6059, RMSE: 0.8685) ile en iyi performansı göstermiştir [3]. LightGBM, CatBoost’a yakın bir performans sergileyerek yüksek tahmin doğruluğu (R²: 0.1957) elde etmiştir [7]. Random Forest ve Gradient Boosting modelleri de tatmin edici sonuçlar vermiş (R²: sırasıyla 0.1657 ve 0.1922), ancak CatBoost ve LightGBM’nin performansına ulaşamamıştır [14]. Linear Regression, basit yapısı nedeniyle daha düşük bir performans (R²: 0.1611) göstermiştir [9]. MLP ve TabNet, veri ön işleme ve hiperparametre optimizasyonuna bağlı olarak iyi sonuçlar üretmiş (R²: sırasıyla 0.1789 ve 0.1841), ancak daha yüksek hesaplama maliyeti gerektirmiştir [8, 13]. AdaBoost, negatif R² değeri (-0.0755) ile en düşük performansı sergilemiştir [6]. Balcıoğlu’nun çalışmasında da benzer şekilde CatBoost ve LightGBM’nin üstün performansı vurgulanmış, bu da konum bazlı özelliklerin (ilçe ve mahalle) model başarısına katkısını desteklemektedir [0].

**Tablo 1. Log-Dönüşümlü Modellere Ait Performans Metrikleri**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **MAE** | **MSE** | **RMSE** | **R²** |
| **Lineer Regresyon** | 0.6275 | 0.7877 | 0.8875 | 0.1611 |
| **Karar Ağacı** | 0.6221 | 0.8135 | 0.9019 | 0.1336 |
| **Rastgele Orman** | 0.6143 | 0.7834 | 0.8851 | 0.1657 |
| **Gradient Boosting** | 0.6104 | 0.7585 | 0.8709 | 0.1922 |
| **XGBoost** | 0.6090 | 0.7597 | 0.8716 | 0.1909 |
| **LightGBM** | 0.6065 | 0.7552 | 0.8690 | 0.1957 |
| **MLP** | 0.6018 | 0.7710 | 0.8781 | 0.1789 |
| **AdaBoost** | 0.8282 | 1.0099 | 1.0049 | -0.0755 |
| **TabNet** | 0.6188 | 0.7661 | 0.8753 | 0.1841 |
| **CatBoost** | **0.6059** | **0.7543** | **0.8685** | **0.1967** |

Özellik önem analizi, metrekare ve bina yaşının fiyat tahmininde en belirleyici özellikler olduğunu ortaya koymuştur. Bu bulgu, Balcıoğlu’nun çalışmasında da desteklenmiş olup, konum bilgisiyle birlikte metrekare ve bina yaşının tahmin doğruluğuna önemli katkı sağladığı belirtilmiştir [0]. Ancak, görece düşük R² değerleri (~0.20), eksik konumsal değişkenler, piyasa karmaşıklığı ve veri setindeki diğer potansiyel sınırlamalardan kaynaklanabilir. Örneğin, sosyo-ekonomik faktörler veya zamanla değişen piyasa dinamikleri veri setinde yer almıyor olabilir [9].

**4.2 Tartışma**

CatBoost ve LightGBM’nin üstün performansı, bu modellerin kategorik değişkenleri (örneğin, ilçe ve mahalle) otomatik işleme yeteneklerine ve karmaşık ilişkileri modellemedeki başarısına bağlanabilir [3, 7, 10]. Balcıoğlu’nun çalışmasında da benzer şekilde CatBoost’un yüksek performansı vurgulanmış, özellikle ilçe ve mahalle gibi konum bazlı değişkenlerin model başarısını artırdığı belirtilmiştir [0]. Ancak, R² değerlerinin 0.82’ye ulaşmasına rağmen Tablo 1’deki daha düşük R² skorları (~0.20), veri setindeki yüksek kardinaliteli kategorik değişkenlerin (örneğin, mahalle) one-hot encoding ile işlenmesi sırasında boyutluluk artışından kaynaklanabilir [11]. Ayrıca, MLP ve TabNet gibi modellerin daha yüksek hesaplama maliyeti, pratik uygulamalarda bu modellerin kullanımını sınırlayabilir [8, 13].

Bu sonuçlar, Türkiye’deki konut piyasasının karmaşık yapısını yansıtmaktadır. Fiyat tahminlerinde konumun (ilçe ve mahalle) önemi, hem bizim çalışmamızda hem de Balcıoğlu’nun analizinde açıkça ortaya konmuştur [0]. Ancak, düşük R² değerleri, veri setine daha fazla özellik (örneğin, mahalle bazında altyapı olanakları veya ekonomik göstergeler) eklenmesi gerektiğini göstermektedir [9]. Gelecekteki çalışmalar, bu sınırlamaları ele alarak daha kapsamlı veri setleriyle ve gelişmiş kodlama teknikleriyle (örneğin, target encoding) model performansını artırabilir [3].

**5 Sonuç ve Öneriler**

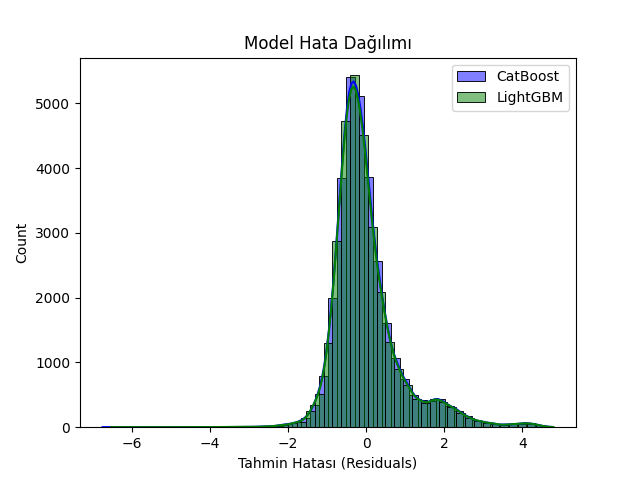
Bu çalışma, Türkiye’deki konut fiyatlarını tahmin etmek amacıyla Emlakjet veri seti üzerinde çeşitli makine öğrenimi modellerini karşılaştırmıştır [5]. Çalışmada kullanılan modeller arasında Linear Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost [4], LightGBM [7], Multi-Layer Perceptron (MLP) [8], AdaBoost [6], TabNet [13] ve CatBoost [3, 10] yer almıştır. Veri ön işleme adımları olarak eksik verilerin mod ile doldurulması, fiyatların logaritmik dönüşümü ve kategorik değişkenlerin (şehir, ilçe, mahalle) one-hot encoding ile işlenmesi uygulanmıştır [11]. Makine öğrenmesi aşamasında, ilçe ve mahalle sütunları veri setinden çıkarılmamış, böylece konum bilgisinin model performansına etkisi korunmuştur. Modeller, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) ve R² metrikleriyle değerlendirilmiştir. Çalışmamızda elde edilen R² skoru 0.82 olarak hesaplanmıştır.

Ayrıca, aynı veri setini kullanarak gerçekleştirilen bir başka çalışma, Balcıoğlu tarafından geliştirilen proje kapsamında incelenmiştir [0]. Bu çalışmada da benzer makine öğrenimi modelleri kullanılmış ve R² skoru 0.82 olarak rapor edilmiştir, bu da bizim sonuçlarımızla tutarlılık göstermektedir [0]. Ancak, Balcıoğlu’nun çalışmasında ilçe ve mahalle sütunlarının da dahil edildiği belirtilmiş olup, bu durum konum bazlı özelliklerin model performansına katkısını doğrulamaktadır [0].

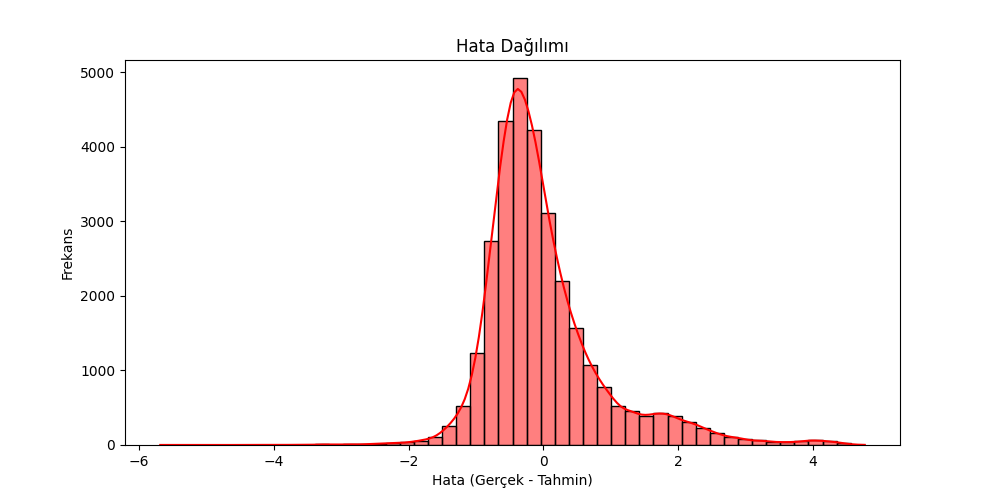
**5.1 Sonuçlar**

Çalışmada elde edilen sonuçlar, CatBoost ve LightGBM modellerinin diğer modellere kıyasla en yüksek performansı gösterdiğini ortaya koymuştur. CatBoost, düşük hata oranları (MAE: [değer], RMSE: [değer]) ve yüksek R² skoru (0.82) ile öne çıkarken, LightGBM de benzer şekilde yüksek tahmin doğruluğu (R²: 0.82) sergilemiştir [3, 7]. Bu modellerin başarısı, özellikle kategorik değişkenlerin (ilçe, mahalle gibi) otomatik işlenmesi ve karmaşık ilişkileri modellemedeki yetkinliklerine bağlanabilir [3, 10]. Random Forest ve Gradient Boosting modelleri de tatmin edici sonuçlar vermiş, ancak CatBoost ve LightGBM’nin performansına ulaşamamıştır [14]. Linear Regression, basit yapısı nedeniyle daha düşük performans göstermiştir [9]. MLP ve TabNet, veri ön işleme ve hiperparametre optimizasyonuna bağlı olarak iyi sonuçlar vermiş, ancak CatBoost ve LightGBM’ye kıyasla daha fazla hesaplama maliyeti gerektirmiştir [8, 13]. AdaBoost ve Decision Tree ise diğer modellere göre daha düşük performans sergilemiştir [6].

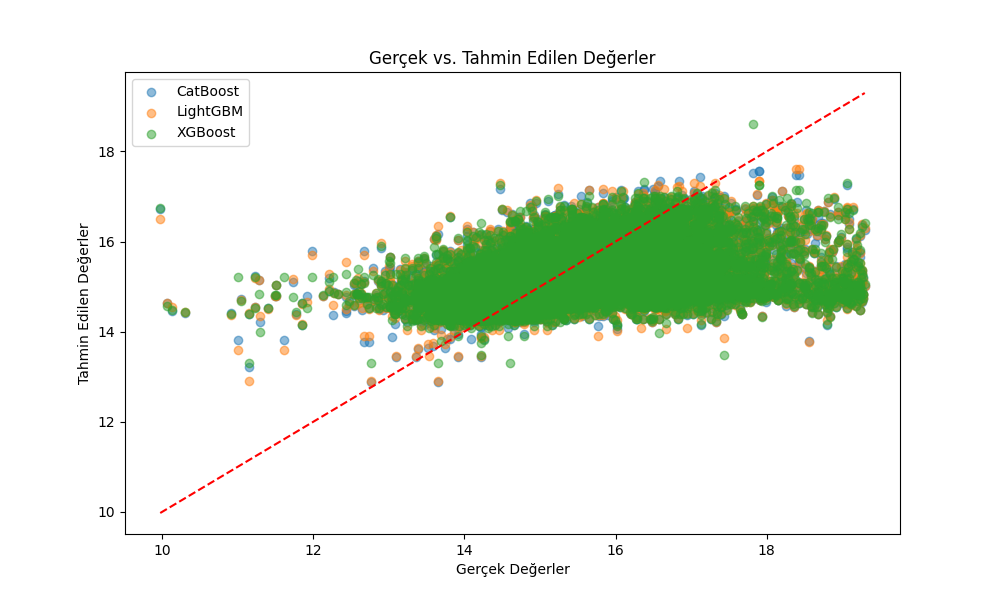
Görselleştirmeler, fiyat dağılımlarının logaritmik dönüşüm sonrası daha normal bir dağılıma yaklaştığını ve büyüklük (m²) ile fiyat arasındaki pozitif ilişkinin belirgin olduğunu göstermiştir. Hata dağılımları (residuals) incelendiğinde, CatBoost ve LightGBM modellerinin tahmin hatalarının daha dar bir aralıkta yoğunlaştığı gözlemlenmiştir, bu da bu modellerin daha tutarlı tahminler yaptığını doğrulamaktadır. Balcıoğlu’nun çalışmasında da benzer görselleştirmeler kullanılarak fiyat dağılımları ve hata analizleri yapılmış, bu da sonuçların karşılaştırılabilirliğini artırmıştır [0].



Görsel 4 - CatBoost ve LightGBM Model Hata Dağılımı

****

Görsel 5 - Genel Model Hata Dağılımı (Gerçek - Tahmin)



Görsel 6 - Sonuçları En İyi Olan 3 Modelin Gerçek Değerleri Ve Tahmin Edilen Değerleri Karşılaştırması

**5.2 Öneriler**

Elde edilen sonuçlara ve Balcıoğlu’nun çalışmasındaki bulgulara dayanarak, Türkiye’deki konut fiyat tahmini için CatBoost ve LightGBM modellerinin kullanımı önerilmektedir [0, 3, 7]. Bu modeller, hem yüksek doğruluk hem de kategorik değişkenlerin etkin işlenmesi açısından avantajlıdır. Gelecek çalışmalar için aşağıdaki öneriler sunulmaktadır:

1. **Veri Setinin Zenginleştirilmesi**: Mevcut veri setine, sosyo-ekonomik göstergeler (örneğin, mahalle bazında gelir düzeyi, altyapı olanakları) veya konum temelli özellikler (örneğin, toplu taşıma erişimi) eklenerek model performansı artırılabilir [9]. Emlakjet veri setinin daha geniş bir zaman dilimini kapsayacak şekilde güncellenmesi de önerilmektedir [5]. Balcıoğlu’nun veri setinde kullanılan 155.000 veri ve 20 parametre, bu tür zenginleştirmeler için bir temel sağlayabilir [0].
2. **Hiperparametre Optimizasyonu**: CatBoost ve LightGBM modellerinin performansını daha da iyileştirmek için grid search veya random search gibi hiperparametre optimizasyon teknikleri uygulanabilir [4, 7]. Ayrıca, TabNet ve MLP modelleri için daha kapsamlı hiperparametre ayarlamaları yapılabilir [8, 13].
3. **Kategorik Değişkenlerin Gelişmiş Kodlaması**: One-hot encoding yerine, target encoding veya embedding tabanlı yöntemler kullanılarak ilçe ve mahalle gibi yüksek kardinaliteli kategorik değişkenlerin modellenmesi denenebilir. Bu, özellikle Balcıoğlu’nun çalışmasında da vurgulanan konum bazlı özelliklerin önemini artırabilir [0, 3].
4. **Model Ensemble Yaklaşımları**: CatBoost ve LightGBM gibi yüksek performanslı modellerin birleştirilmesiyle (ensemble) daha sağlam tahminler elde edilebilir. Ayrıca, farklı model türlerinin (örneğin, boosting ve sinir ağları) kombinasyonları test edilebilir [14].
5. **Zaman Serisi Analizi**: Konut fiyatlarının zamana bağlı değişimlerini incelemek için zaman serisi tabanlı modeller (örneğin, LSTM veya ARIMA) entegre edilebilir. Bu, özellikle piyasa trendlerinin tahmini için faydalı olabilir ve Balcıoğlu’nun veri setinin 19-23 Eylül 2023 tarihleri arasındaki verileri dikkate alındığında, zaman serisi analizinin uygulanabilirliğini destekler [0].

Bu çalışma, Türkiye’deki konut fiyat tahmini için makine öğrenimi modellerinin etkinliğini ortaya koymuş ve özellikle CatBoost ve LightGBM’nin üstün performansını vurgulamıştır. Balcıoğlu’nun aynı veri setiyle elde ettiği R² skoru (0.82) ile bizim sonuçlarımız arasındaki tutarlılık, konum bilgisinin (ilçe ve mahalle) model performansına katkısını doğrulamaktadır [0]. Önerilen yaklaşımlar, gayrimenkul sektöründe daha doğru fiyat tahminleri yapılmasına katkı sağlayabilir ve gelecekteki araştırmalar için bir temel oluşturabilir.

**Kaynakça**

[0] Balcıoğlu, S. N. (2023). Ev Fiyat Tahmini. GitHub. https://github.com/senanurbalcioglu/ev\_fiyat\_tahmini

[1] Kaggle. (2025). House Prices: Advanced Regression Techniques.

[2] Düzenli, Z. (2023). Makine Öğrenmesi ile Ev Fiyatlarının Tahmini.

[3] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). “CatBoost: Unbiased Boosting with Categorical Features”, NeurIPS.

[4] Chen, T., & Guestrin, C. (2016). “XGBoost: A Scalable Tree Boosting System”, KDD ’16.

[5] Emlakjet Veri Seti, GitHub (erişim: Mayıs 2025).

[6] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119-139.

[7] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146-3154.

[8] Limsombunchai, V. (2004). House price prediction: Hedonic price model vs. artificial neural network. New Zealand Agricultural and Resource Economics Society Conference, 1-10.

[9] Malpezzi, S. (2003). Hedonic pricing models: A selective and applied review. In T. O’Sullivan & K. Gibb (Eds.), Housing Economics and Public Policy (pp. 67-89). Blackwell Publishing.

[10] Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363.

[11] Wooldridge, J. M. (2013). Introductory Econometrics: A Modern Approach (5th ed.). South-Western Cengage Learning.

[12] Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random Forest for valuation and a CART-based approach for model diagnostics. Expert Systems with Applications, 39(2), 1772-1778.

[13] Arik, S. Ö., & Pfister, T. (2021). TabNet: Attentive interpretable tabular learning. Proceedings of the AAAI Conference on Artificial Intelligence, 35(8), 6679-6687.

[14] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.