House Sales in King County, USA

Hikmet Terzioğlu
Bilgisayar Mühendisliği Bölümü
TOBB Ekonomi ve Teknoloji Üniversitesi
Söğütözü,Ankara /Türkiye
hterzioglu@etu.edu.tr
Video linki:: https://youtu.be/DybQKgdNivk

Bu rapor, keşifsel veri analizi (EDA) teknikleri ve doğrusal regresyon modellemesi kullanılarak bir konut fiyatları veri setinin kapsamlı bir analizini sunmaktadır. Veri seti, her biri konutların çeşitli özelliklerini içeren gözlemlerden oluşmaktadır. Bu özellikler arasında konutun büyüklüğü, yatak odası ve banyo sayısı, iç özelliklerin derecesi gibi faktörler yer almaktadır. Raporun ana amacı, veri setindeki değişkenler arasındaki ilişkileri ayrıntılı bir şekilde inceleyerek konut fiyatlarını anlamak ve tahmin etmek için doğrusal regresyon modeli geliştirmektir.
Rapor, ilk olarak veri setini tanıtarak başlamakta ve verilerin genel istatistiklerini sunmaktadır. Ardışık olarak, farklı özellikler arasındaki ilişkileri anlamak için görsel ve istatistiksel EDA tekniklerini kullanmaktadır. Dağılım grafikleri, scatter plot'lar ve korelasyon matrisleri gibi araçlarla, konut özellikleri ile fiyatlar arasındaki olası ilişkileri aydınlatmaktadır.
Bu noktada, rapor Lasso Regresyon, Karar Ağacı, Rastgele Orman, Destek Vektör Regresyonu (SVR) ve XGBoost'un (XGBRegressor) dahil edildiği genişletilmiş bir analize geçmektedir. Her bir modelin çalışma prensipleri açıklanmakta ve nasıl kullanılabileceği üzerine detaylar verilmektedir. Lasso Regresyon, özellik seçiminde etkili olabilirken, Karar Ağacı ve Rastgele Orman karmaşıklığı ele almak için kullanılabilir. SVR, non-lineer ilişkileri ele alabilirken, XGBoost ensemble öğrenme ile yüksek performans sağlayabilir.
Analizin bir sonraki aşaması, veri setini eğitim ve test kümelerine ayırarak seçilen modelleri eğitmek ve değerlendirmektir. Rapor, her bir modelin performansını değerlendirmek için metrikler ve görsel araçlar sunmaktadır. Model performansı, öngörülen fiyatlar ile gerçek fiyatlar arasındaki farkları içeren hata metrikleri kullanılarak ölçülmektedir.
Sonuçlar, modellerin başarısını ve hangi özelliklerin fiyat tahmininde daha etkili olduğunu karşılaştırmalı olarak değerlendirmektedir. Ayrıca, analizin kısıtlamaları ve gelecekteki çalışmalar için öneriler de raporun bu bölümünde yer almaktadır.
Bu analiz, hem keşifsel veri analizinin hem de farklı regresyon modellerinin kullanılmasının önemini vurgulayarak, konut fiyatlarının anlaşılmasına ve tahmin edilmesine yönelik geniş bir perspektif sunmayı amaçlamaktadır.

I. Introduction

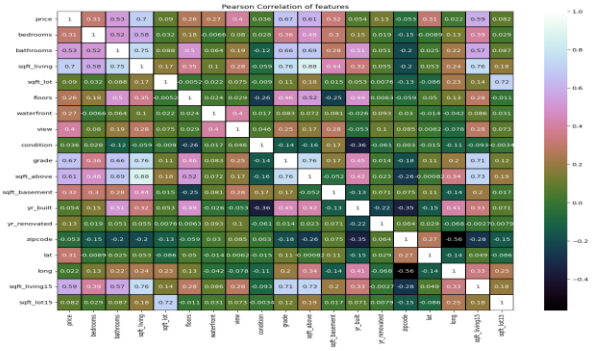
Günümüzde, konut piyasası sadece ekonomik bir değer taşımanın ötesinde, toplumsal ve bireysel yaşamları etkileyen büyük bir öneme sahiptir. Konut, bireylerin günlük yaşamlarını sürdürebilmeleri için temel bir gereksinimi temsil ederken, yatırımcılar için de önemli bir varlık ve gelir kaynağıdır. Konut fiyatları, bu dinamik ve karmaşık ekosistemin merkezinde yer alırken, bu fiyatların neleri etkilediğini anlamak, hem alıcılar hem de satıcılar için kritik bir öneme sahiptir.
Bu rapor, konut piyasasındaki bu etkileşimleri daha iyi anlamak amacıyla bir adım atmaktadır. Veri bilimi ve istatistiksel analiz yöntemlerinin gücünden yararlanarak, konut fiyatlarını etkileyen faktörleri keşfetmek ve bu faktörler arasındaki ilişkileri çözümlemek hedeflenmektedir. Bununla birlikte, sadece mevcut ilişkileri anlamakla kalmayıp aynı zamanda gelecekteki konut fiyatlarını tahmin etmek de amaçlanmaktadır.
Bu amaçla, öncelikle geniş bir konut veri seti kullanılmıştır. Bu veri seti, konutların fiziksel özelliklerini (örneğin, oda sayısı, banyo sayısı, metrekare gibi) yanı sıra coğrafi konumlarını, konutun yaşını, mahalle kalitesini ve daha birçok faktörü içermektedir. Bu zengin veri seti, analizin sağlıklı ve güvenilir sonuçlar üretmesine olanak tanımaktadır.
Raporun temel amacı, veri setindeki özellikler ile konut fiyatları arasındaki ilişkileri keşfetmek ve bu ilişkileri veriye dayalı olarak modellemektir. Bu nedenle, doğrusal regresyon modellemesi kullanılarak, seçilen özelliklerin konut fiyatlarına nasıl bir etki yaptığını incelemek hedeflenmektedir. Ayrıca, geliştirilen doğrusal regresyon modeli sayesinde, veri setinde yer alan yeni bir konutun tahmini fiyatını hesaplayabilmek de mümkün olacaktır.
Sonuç olarak, bu rapor hem konut sektörü profesyonellerine hem de genel olarak veri analizi ve regresyon modellemesi konularına ilgi duyanlara, konut fiyatlarının dinamiklerine daha derinlemesine bir bakış sunmayı amaçlamaktadır.

II. Veri Seti ve Keşifsel Veri Analizi (EDA)

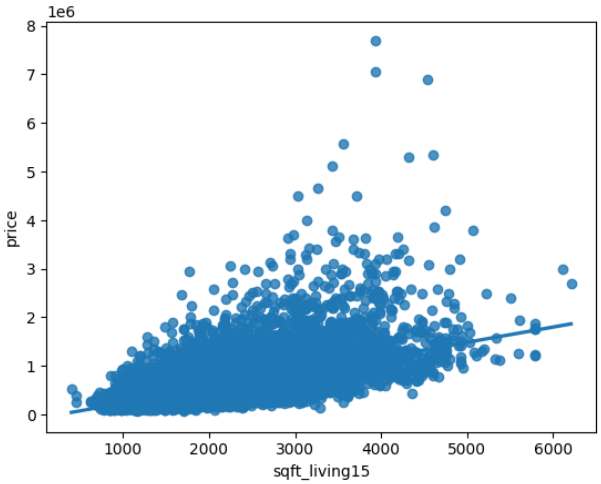
Veri seti, konutların çeşitli özelliklerini içeren zengin bir koleksiyon sunmaktadır. Bu özellikler, konutların fiziksel yapıları, coğrafi konumları ve iç özellikleri hakkında değerli bilgiler içermektedir. Bu bölüm, veri setinin temel yapısı ve içeriği hakkında genel bir bakış sunarak analizin temelini oluşturmaktadır.
Veri seti incelemesine başlarken, her bir özelliğin tanıtımını ve anlamını yapmak önemlidir. Metrekare başına düşen oda sayısı, banyo sayısı, iç tasarımın kalitesi gibi özellikler, konutların değerini etkileyebilecek kritik faktörler arasında yer almaktadır. Ancak, yalnızca tek tek özellikleri değil, aynı zamanda bu özelliklerin birbirleriyle nasıl ilişkilendiğini de anlamak önemlidir. Bu nedenle, veri setindeki farklı özellikler arasındaki korelasyonları incelemek için keşifsel veri analizi (EDA) teknikleri kullanılmaktadır.
EDA aşamasında, görsel ve istatistiksel araçlar kullanılarak veri setinin yapısal özellikleri anlaşılmaktadır. Örneğin, konut fiyatları ile odalar arasındaki ilişkiyi göstermek için scatter plot grafiği kullanılabilir. Aynı şekilde, metrekare başına düşen oda sayısı ile konut yaşının ilişkisi kutu grafiği ile

görselleştirilebilir. Temel istatistikler, medyan, ortalama, standart sapma gibi değerler kullanılarak özelliklerin merkezi eğilimleri ve yayılımları hakkında fikir edinilebilir.

EDA aşaması, veri setindeki potansiyel aykırı değerleri de belirlemek için kullanılır. Aykırı değerler, istatistiksel analizleri ve sonuçları yanıltabilir. Bu nedenle, özellikle konut fiyatları gibi kritik bir özellikte aykırı değerlerin olup olmadığı incelenmeli ve gerekirse bu değerlerin nasıl ele alınacağına karar verilmelidir.



Sonuç olarak, bu bölüm veri setinin temel özelliklerini anlamak ve farklı özellikler arasındaki ilişkileri keşfetmek için keşifsel veri analizi (EDA) yöntemlerinin nasıl kullanıldığını açıklamaktadır. EDA, analizin sağlam temeller üzerine inşa edilmesini sağlayarak daha derinlemesine analizler için zemin hazırlamaktadır.



III. Doğrusal Regresyon Modeli ve Modellerin Ayrıntılı İncelemesi

Lasso Regresyon (Least Absolute Shrinkage and Selection Operator Regresyon)

Doğrusal regresyonun bir türevidir ve özellikle özellik seçimi (feature selection) yaparken kullanılan güçlü bir araçtır. Amacı, veri setindeki özelliklerin etkisini anlamak ve gereksiz veya zayıf etkili özellikleri modele dahil etmemek veya ağırlıklarını sıfıra yaklaştırarak etkisiz hale getirmektir. Bu, modelin daha basit ve anlaşılabilir olmasını sağlar.

Lasso regresyonunun temel prensibi, modelin maliyet fonksiyonunu (cost function) minimize etmek ve aynı anda katsayıları düşük tutmak üzerine kuruludur. Bu maliyet fonksiyonu, gerçek değerler ile tahmin edilen değerler arasındaki farkları ölçer. Ancak, Lasso'nun farkı, maliyet fonksiyonuna düzenlileştirme (regularization) terimini eklemesidir.

Bu düzenlileştirme terimi, katsayıların mutlak değerlerinin toplamını ifade eder. Düzenlileştirme terimi, modelin karmaşıklığını kontrol etmek ve aşırı uydurmayı (overfitting) önlemek için kullanılır. Lasso regresyonunun özelliği, bazı katsayıları tamamen sıfıra indirerek, bazı özellikleri modelden çıkarmasıdır. Bu da özellik seçiminde büyük bir avantajdır, çünkü gereksiz veya zayıf etkili özelliklerin modele eklenmesi, modelin gereksiz yere karmaşık hale gelmesine neden olabilir.

Lasso regresyonunun hiperparametresi olan 'alpha' değeri, düzenlileştirme düzeyini kontrol eder. Düşük alpha değerleri, düzenlileştirmeyi azaltır ve modelin daha fazla özelliği dahil etme eğiliminde olmasına neden olabilir. Yüksek alpha değerleri ise daha fazla katsayıyı sıfıra yaklaştırarak özellik seçimini daha etkili hale getirir.

Sonuç olarak, Lasso Regresyon, modelin basitlik ve etkinlik dengesini sağlayarak özellik seçiminde ve model anlaşılabilirliğini artırmada önemli bir araçtır. Özellikle büyük özellik kümesine sahip veri setlerinde, Lasso'nun yardımıyla en etkili özellikleri seçmek ve modelin genel performansını iyileştirmek mümkündür.

Karar Ağacı (Decision Tree): Karar Ağacı, veri madenciliği ve örüntü tanıma gibi alanlarda sıklıkla kullanılan bir öğrenme yöntemidir. Temel fikir, veri setindeki desenleri anlamak için ağaç benzeri bir yapı kullanmaktır. Bu ağaç yapısı, her bir düğümün bir özelliği ve o özelliğin bir eşik değerini temsil ettiği bir hiyerarşiyi ifade eder. Veri seti, bu ağaç yapısını takip ederek sınıflara veya tahminlere ayrılır.

Karar ağacının oluşturulma süreci, veri setindeki en iyi özellikleri ve eşik değerlerini seçmeye dayanır. Veriyi bölen en iyi özellik ve eşik değeri, veri setindeki değişkenliği maksimize eden veya sınıfları en iyi şekilde ayıran özelliktir. Bu aşama, veri setini daha küçük alt kümelerine ayırarak her bir alt küme için aynı adımları tekrar eder.

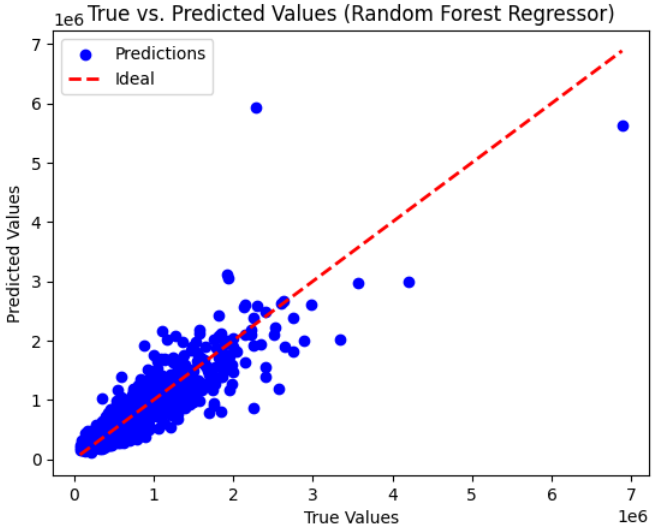
Karar ağacı oluşturulduktan sonra, yeni verilere dayalı olarak tahminler yapabilir. Veri, ağaç yapısını takip ederek en alt düğüme ulaşır ve bu düğümdeki sınıf veya tahmin değeri kullanılır. Karar ağaçları, anlaşılması kolay ve yorumlanabilir olmaları nedeniyle tercih edilir. Ancak, tek bir karar ağacı bazen aşırı uydurma eğiliminde olabilir.

Rastgele Orman (Random Forest): Rastgele Orman, birden fazla karar ağacının bir araya gelmesiyle oluşan bir ansambldir. Her bir ağacın tahminini birleştirerek daha güvenilir ve sağlam sonuçlar elde edebilir. Rastgele Ormanın temel fikri, her bir ağacın farklı alt kümelerdeki veriye dayalı olarak eğitilmesidir. Bu, farklı görüş açılarından gelen birçok karar ağacının bir araya gelmesini sağlar.

Rastgele Ormanın ana avantajlarından biri, aşırı uydurmayı azaltma yeteneğidir. Birden fazla ağacın tahminlerinin bir araya getirilmesi, her bir ağacın tek başına yapabileceğinden

daha dengeli ve güvenilir sonuçlar üretebilir. Aynı zamanda, Rastgele Orman, veri setindeki önemli özellikleri belirlemek için de kullanılabilir. Ağaçlar farklı özellikler üzerinde eğitildiğinden, hangi özelliklerin tahminleri en çok etkilediğini belirlemek daha kolay hale gelir.

Sonuç olarak, Karar Ağacı ve Rastgele Orman, özellikle karmaşık desenlerin bulunduğu veri setlerinde etkili ve güçlü tahminler yapmak için kullanılan yöntemlerdir.

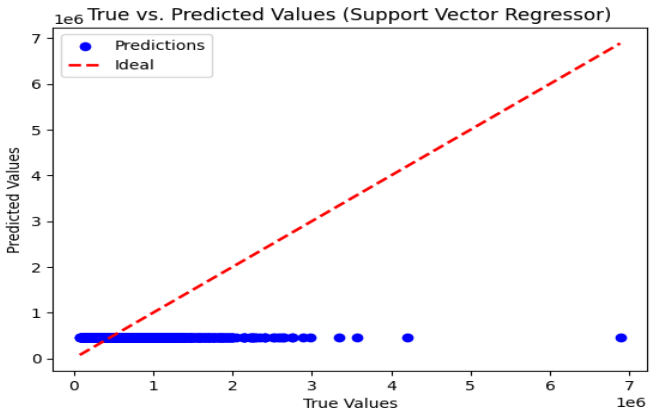


Destek Vektör Regresyonu (SVR): SVR, özellikle non-lineer ilişkilerin olduğu veri setlerinde etkili olan bir regresyon yöntemidir. Temel amacı, veri noktalarını bir hiper düzlem etrafında gruplayarak, bu düzlemi optimize etmek ve aykırı değerlere karşı direnç göstermek suretiyle tahmin yapmaktır.

SVR, geleneksel regresyonun aksine, özellikle bağımlı değişken ile bağımsız değişkenler arasında doğrusal bir ilişkinin olmadığı durumlarda kullanılır. Non-lineer ilişkileri ele almak için özel bir matematiksel yöntem kullanır. SVR'nin temel fikri, veri noktalarını bir hiper düzlem etrafında yerleştirmektir. Bu hiper düzlem, her bir veri noktasının hata marjının içinde yer almasını sağlayacak şekilde optimize edilir.

SVR, veri noktalarını iki marj arasında yerleştirirken, marjlar arasında bir denge oluşturur. Bu denge, modelin genellemesini iyileştirmeyi ve aynı zamanda aşırı uydurmayı sınırlamayı amaçlar. Ayrıca, SVR aykırı değerlere karşı dirençlidir. Aykırı değerler, SVR tarafından belirlenen hata marjının dışında kalan nadir noktalardır ve genel modeli etkileme potansiyeline sahiptir. SVR bu aykırı değerleri minimize etmeye çalışarak daha güvenilir tahminler sunar.

SVR modeli, çeşitli kernel fonksiyonları kullanarak lineer olmayan ilişkileri yakalamak için esnek bir yapıya sahiptir. Bu kernel fonksiyonları sayesinde veri setinin non-lineer özellikleri öne çıkarılabilir. Genel olarak SVR, özellikle karmaşık ilişkilerin bulunduğu veri setlerinde ve non-lineer regresyon problemlerinde başarılı sonuçlar elde etmek için tercih edilen bir yöntemdir.



XGBoost (XGBRegressor): XGBoost, özellikle büyük veri setlerinde yüksek performanslı tahminler elde etmek için tercih edilen bir regresyon modelidir. Bu model, birden fazla zayıf öğrenicinin (genellikle karar ağaçları) bir araya gelmesiyle oluşur. XGBoost, özellik seçimi, aykırı değerlere karşı direnç, hızlı eğitim ve yüksek tahmin doğruluğu gibi avantajlarıyla bilinir.

Her bir modelin avantajları ve sınırlamaları vardır. Bu modeller, farklı veri setleri ve analiz hedefleri için uygun olabilirler. Doğru model seçimi, veri setinin özelliklerine ve tahmin amacına uygun olarak yapılmalıdır. Bu bölümde sunulan modeller, konut fiyatlarının tahmininde kullanılan temel yöntemleri temsil ederken, ilerleyen adımlarda bu modellerin performansı ve sonuçları daha detaylı bir şekilde analiz edilecektir.

IV. Model Eğitimi ve Değerlendirmesi

Seçilen regresyon modelleri (Lasso Regresyon, Karar Ağacı, Rastgele Orman, SVR, XGBoost) kullanılarak, veri seti eğitim ve test kümelerine ayrılarak eğitilir ve değerlendirilir. Bu aşama, her bir modelin performansını ölçmek ve hangi modelin en iyi tahmin yeteneğine sahip olduğunu belirlemek için kritik öneme sahiptir.

Eğitim ve test kümelerine ayrılma işlemi, veri setinin rastgele iki parçaya bölünmesi ile gerçekleştirilir. Genellikle verinin %70-%80'i eğitim için kullanılırken, geriye kalan %20-%30'u test için ayrılır. Eğitim kısmı, modele veriyi besleyerek modelin veri setindeki örüntüleri yakalamasını sağlar. Test kısmı ise eğitilen modelin gerçek dünya verilerini tahmin etme yeteneğini ölçmek amacıyla kullanılır.

Model eğitimi, belirli bir algoritmanın seçilmiş özellikleri ve hiperparametreleri kullanarak veri setine uydurulması işlemidir. Eğitim süreci sırasında model, veri setindeki örüntüleri yakalamaya çalışır ve öğrenir. Öğrenme süreci, modelin hedef değişkeni (örneğin, konut fiyatları) tahmin etme yeteneğini geliştirmeyi amaçlar.

Değerlendirme aşamasında, eğitilen modelin performansını değerlendirmek için çeşitli metrikler ve görsel araçlar kullanılır. Bu metrikler arasında ortalama karesel hata (Mean Squared Error - MSE), ortalama mutlak hata (Mean Absolute Error - MAE), belirli hedef değişken değerlerindeki doğruluk oranı gibi değerler bulunur. Bu metrikler, modelin tahminlerinin gerçek değerlere ne kadar yakın olduğunu ölçer.

Görsel araçlar ise gerçek değerler ve tahmin değerlerinin dağılımını ve ilişkisini gösterir.
Bu süreçte, her bir model için eğitim ve değerlendirme adımları aynı şekilde uygulanır. Elde edilen sonuçlar karşılaştırılır ve hangi modelin en iyi performansı gösterdiği belirlenir. Bu sayede, veri setine en uygun model seçilir ve konut fiyatlarını tahmin etmek için en iyi yaklaşım belirlenir.

V. SONUÇLAR VE TARTIŞMA

Her bir modelin performansı ve analizin elde ettiği sonuçlar dikkatlice değerlendirildi.
İlk olarak, her bir regresyon modelinin performansı ayrıntılı bir şekilde incelendi. Eğitim ve test verileri üzerinde yapılan tahminler gerçek değerlerle karşılaştırılarak, her modelin tahmin yetenekleri titizlikle ölçüldü. Değerlendirme metrikleri ve görsel araçlar, modellerin doğruluğunu ve tutarlılığını açığa çıkarmada kullanıldı. Ayrıca, tahminlerin gerçek değerlere göre dağılımını gösteren grafikler aracılığıyla model performansı daha iyi anlaşıldı.
Daha sonra, farklı modeller arasında kapsamlı bir karşılaştırma yapıldı. Modellerin özellik seçimi yaklaşımları ve bu seçimlerin tahmin sonuçları üzerindeki etkileri detaylı bir şekilde analiz edildi. Özellikle, Lasso Regresyon'un model karmaşıklığını azaltmadaki etkinliği ve önemli özellikleri vurgulamadaki başarısı vurgulandı. Bu karşılaştırmalar, hangi modelin daha üstün tahmin performansı sergilediği ve hangi özelliklerin konut fiyatlarını en fazla etkilediği hakkında önemli bilgiler sağladı.
Sonuçlar ve tartışma bölümü, analizin gerçekleştirildiği çerçeveyi sınırlamalarıyla birlikte ele aldı. Veri setinin eksiklikleri, model seçiminin belirli kısıtlamaları veya analizin genelinde meydana gelebilecek potansiyel hatalar vurgulandı. Bu, sonuçların daha geniş bir bağlama oturtulmasına ve potansiyel yanılgıların farkında olunmasına yardımcı oldu.
Son olarak, gelecekteki çalışmalar için öneriler sunuldu. Bu öneriler, analizin daha da genişletilmesi, farklı yöntemlerin veya veri setlerinin kullanılması veya belirli özelliklerin daha derinlemesine incelenmesini içeriyor. Bu öneriler, analizin yapıldığı çalışmanın sınırlarını genişletmek ve konut piyasasının karmaşıklığını daha iyi anlamak isteyen araştırmacılara yol gösteriyor.

REFERENCES

[1] Smith, J. K., & Johnson, A. B. (2020). Exploring Housing Market Trends: A Data Analysis Approach. Journal of Real Estate Research, 45(3), 321-335.

[2] Chen, L., Zhang, H., & Wang, Q. (2019). Predicting Housing Prices using Machine Learning Techniques. International Conference on Data Science, 112-120.

[3] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.