

Open in app ↗



Search

Get unlimited access to the best of Medium for less than \$1/week. [Become a member](#)

Makine Öğrenmesi Dersleri 5b: Random Forest (Regresyon)

Hakkı Kaan Simsek · [Follow](#)

Published in Veri Bilimi Türkiye

4 min read · Sep 1, 2018



Share



More



Rassal orman (Random Forest), hiper parametre kestirimi yapılmadan da iyi sonuçlar vermesi hem regresyon hem de sınıflandırma problemlerine uygulanabilir olmasından dolayı popüler makine öğrenmesi modellerinden biri. Rassal ormanı anlamak için önce bu modelin temel blogu olan karar ağaçlarını anlamak gerekiyor. 3. dersi bu konuya ayırmıştık kısaca başarılı bir karar ağacı günlük hayatta bilgi

kazancını arttıracak doğru sorular soran ve isabetli tahminler yapan insanlara benzetilebilir.

Fakat geleneksel yöntemlerden biri olan karar ağaçlarının en büyük problemlerinden biri aşırı öğrenme-veriyi ezberlemedir (overfitting). Rassal orman modeli bu problemi çözmek için hem veri setinden hem de öznitelik setinden rassal olarak 10'larca 100'lerce farklı alt-setler seçiyor ve bunları eğitiyor. Bu yöntemle 100'lerce karar ağacı oluşturuluyor ve her bir karar ağacı bireysel olarak tahminde bulunuyor. Günün sonunda problemimiz regresyonsa karar ağaçlarının tahminlerinin ortalamasını problemimiz sınıflandırmaysa tahminler arasında en çok oy alanı seçiyoruz.

Şimdi bütün bunları basit bir örnekle açıklamaya çalışalım: Örneğin bu akşam güzel bir film izlemek istiyorsunuz ve kafanız karışık. Bir arkadaşınızı ararsanız ve o size tercih ettiğiniz film türü, süre, yıl, oyuncu-yönetmen, hollywood-alternatif vs. soru setinden çeşitli sorularla daha önce izlediğiniz filmlere (training set) göre bir tahminde bulunursa bu karar ağacı olur. Eğer 20 arkadaşınız bu soru setinden farklı sorular seçip verdiğiniz cevaplara göre tavsiyede bulunursa ve siz en çok tavsiye edilen filmi seçerseniz bu rassal orman olur.

Rassal orman modelinde farklı veri setleri üzerinde eğitim gerçekleştiği için varyans, diğer bir deyişle karar ağaçlarının en büyük problemlerinden olan overfitting azalır. Ayrıca bootstrap yöntemiyle oluşturduğumuz alt-veri kümelerinde outlier bulunma şansını da düşürmüş oluruz.

Random forest modelinin diğer bir özelliği bize özniteliklerin ne kadar önemli olduğunu vermesi. (Bir özniteliğin önemli olması demek o özniteliğin bağımlı değişkendeki varyansın açıklanmasına ne kadar katkı yaptığıyla alakalı.) Random forest algoritmasına sayıda öznitelik verip en faydalı x tanesini seçmesini isteyebiliriz ve istersek bu bilgiyi istediğimiz başka bir modelde kullanabiliriz.

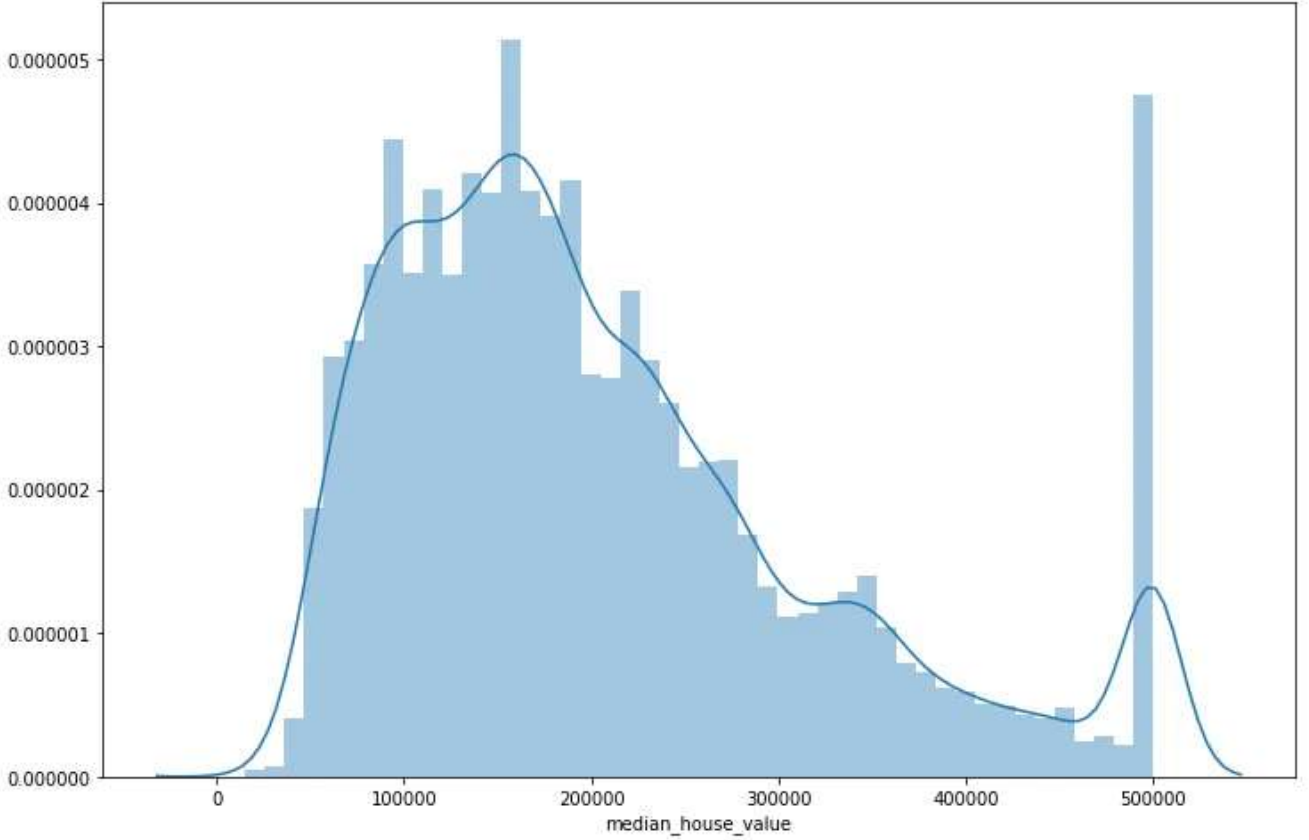
İlk derslerde olduğu gibi bu dersi takip etmek için de bilgisayarınıza Python kurmanıza veya veri setini indirmenize gerek yok. İzlemeniz gereken adımlar sırasıyla şöyle:

- Google hesabınızı açın.
 - <https://colab.research.google.com/> adresine gidin.
 - NEW PYTHON 3 NOTEBOOK'a tıklayın.
 - Oradaki satıra aşağıdaki kodu yapıştırın ve play tuşuna basın.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
longitude      20640 non-null float64
latitude       20640 non-null float64
housing_median_age  20640 non-null float64
total_rooms    20640 non-null float64
total_bedrooms 20433 non-null float64
population     20640 non-null float64
households     20640 non-null float64
median_income  20640 non-null float64
median_house_value 20640 non-null float64
ocean_proximity 20640 non-null object
dtypes: float64(9), object(1)
memory usage: 1.6+ MB
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY

Modelimizdeki hedef değişken ‘median_house_value’ olduğu için onun dağılımına bakıyoruz ve veri setinin geri kalanından ayırıyoruz.



Makine öğrenmesi modelleri kategorik değişkenleri algılayamadığı için 'object' tipindeki değişkenleri one-hot-encoding yöntemiyle 0 ve 1'lere ayırıyoruz. Ayrıca 'total_bedroom' değişkenindeki eksik değerleri doldurmamız gerekiyor.

Yapmamız gerekenler sırasıyla şöyle:

- ocean_proximity değişkeni `pd.get_dummies()` fonksiyonuyla zenginleştirilir.
- Bu değişken veri setinden atılır.
- total_bedroom değişkeni yine o sütunun `median()` değeriyle doldurulur.

Modelimizi kurmaya hazırız şimdi standart modelleme süreçlerini uygulayacağız.

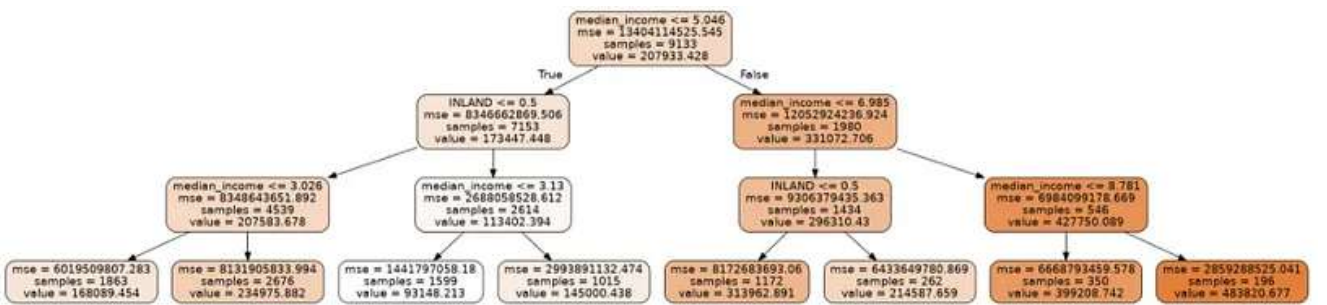
1. Veriyi eğitim ve test alt-veri setlerine ayırma.
2. Karar ağacı modeli oluşturma.
3. Modeli eğitim verisine ‘fit’ etme.

```
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=3,  
                        max_features='auto', max_leaf_nodes=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=1,  
                        oob_score=False, random_state=42, verbose=0, warm_start=False)
```

4. Görmediğimiz test verisine modele verip tahminde bulunma.
5. Gerçek değerle tahmin arasındaki benzerliğe göre mean absolute error, mean squared error ve root mean squared error hesaplama.

```
Mean Absolute Error (MAE): 54297.29  
Mean Squared Error (MSE): 5550290067.41  
Root Mean Squared Error (RMSE): 74500.27
```

6. Rastgele ormandan bir karar ağacı çekip görselleştirme.

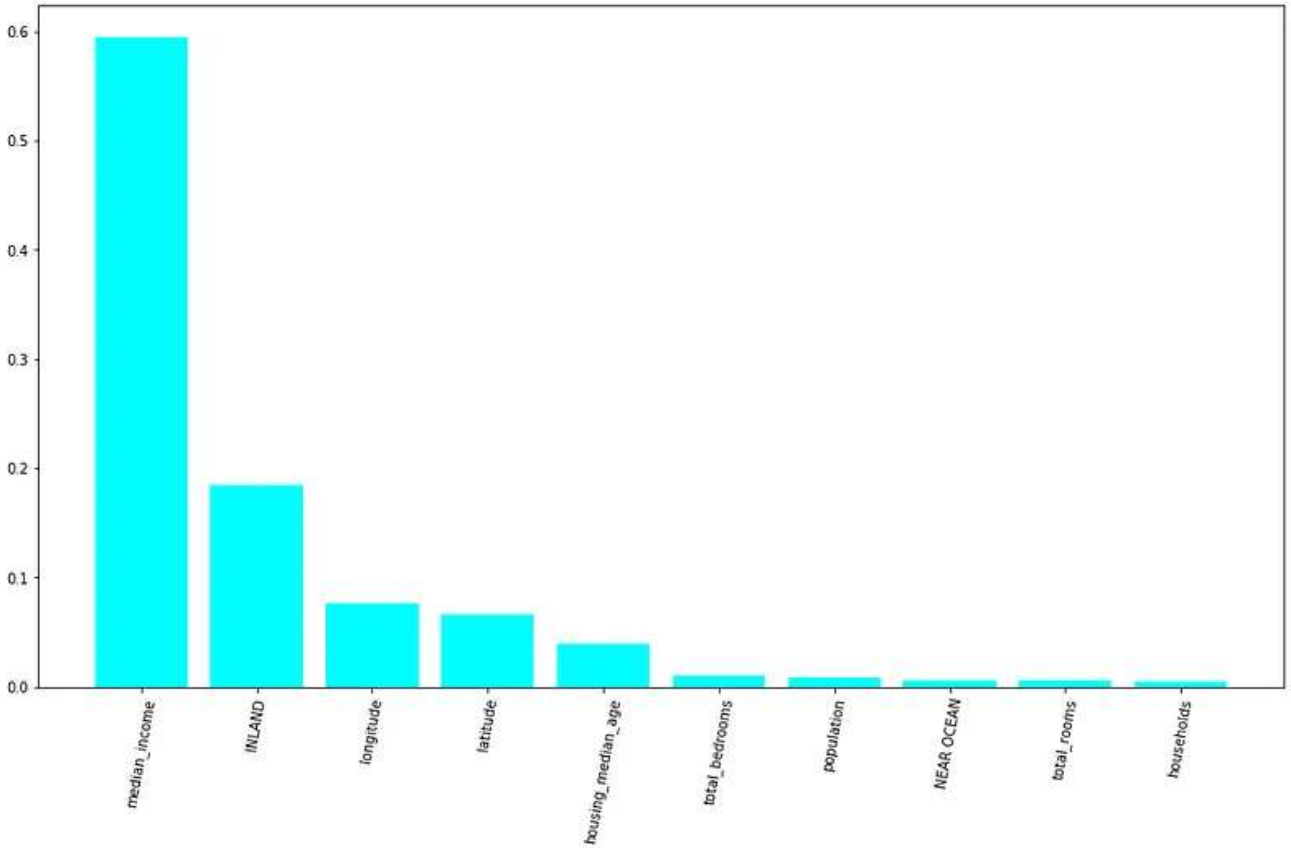


7. Sınıflandırma modeli kurulurken yapılan öznitelik önem sıralamasını görselleştirme.

Burada derinliği 8 olan yeni bir model kurup o modelin başarı oranına bakıyoruz ve öznitelik sıralamasını görselleştiriyoruz.

Mean Absolute Error (MAE): 39691.24
Mean Squared Error (MSE): 3307703633.87
Root Mean Squared Error (RMSE): 57512.64

Feature importances based on Random Forest Regressor



Başarı oranımız ciddi bir şekilde arttı. Bu çalışma standart bir veri bilimi projesinin en basit hali olarak düşünülebilir.

- Veri tiplerini kontrol etme/düzeltilme.
- Açıklayıcı veri analizi ve görselleştirme.
- Eksik verileri tahmin etme/veri atama.
- Kategori tipindeki verileri one-hot encoding ile nümerik formata çevirme.
- Veri setini eğitim ve test veri-setlerine ayırma.
- Modeli eğitme ve test verisi üzerinde tahmin yapma.
- Sınıflandırma başarı metriklerine bakma.
- Karar ağacını görselleştirme.
- Yeni modelin başarı metriklerine bakma ve öznitelik sıralamasını görselleştirme.

Çalışmadaki veri setine ve kodlara [şuradan](#) ulaşabilirsiniz.

Sorunuz olursa bana [LinkedIn](#) veya [Twitter](#) hesaplarından yazabilirsiniz.

[Makine Öğrenmesi](#)[Yapay Zeka](#)[Yapay Öğrenme](#)[Büyük Veri](#)[Buyuk Veri](#)[Follow](#)

Written by Hakkı Kaan Simsek

2.1K Followers · Editor for Veri Bilimi Türkiye

Head of Data @scantrust | AWS Solution Architect <https://github.com/kaan-simsek>

More from Hakkı Kaan Simsek and Veri Bilimi Türkiye



Hakkı Kaan Simsek in Veri Bilimi Türkiye

Makine Öğrenmesi Dersleri 6: NLP'ye Giriş

NLP yani Doğal Dil İşleme, doğal dillerin kurallı yapısının çözümlenerek anlaşılması veya yeniden üretilmesi amacını taşır. Bu çözümlemenin...

4 min read · Apr 8, 2018



247



2



Hakkı Kaan Simsek in Veri Bilimi Türkiye

Makine Öğrenmesi Dersleri 10: Sınıflandırma Modellerinde Başarı Kriterleri

Sınıflandırma algoritmalarını kullanarak yapılan çalışmalarda en büyük yanılgılardan biri başarı kriteri olarak sadece doğruluk oranına...

4 min read · May 12, 2018



632



1



Hakkı Kaan Simsek in Veri Bilimi Türkiye

Makine Öğrenmesi Dersleri 4a: Lojistik Regresyon

Gözetimli makine öğrenmesi (supervised machine learning) ve istatistik modelleri temel olarak iki problemi çözmeye çalışır:

4 min read · Mar 6, 2018



303



3





Hakkı Kaan Simsek in Veri Bilimi Türkiye

Makine Öğrenmesi Dersleri 7: Boyut Azaltma (Dimensionality Reduction)

Boyut Azaltma (Dimensionality Reduction) veri bilimi için oldukça önemli bir yöntem. Başlıca sebepleri şöyle:

3 min read · Apr 14, 2018



183



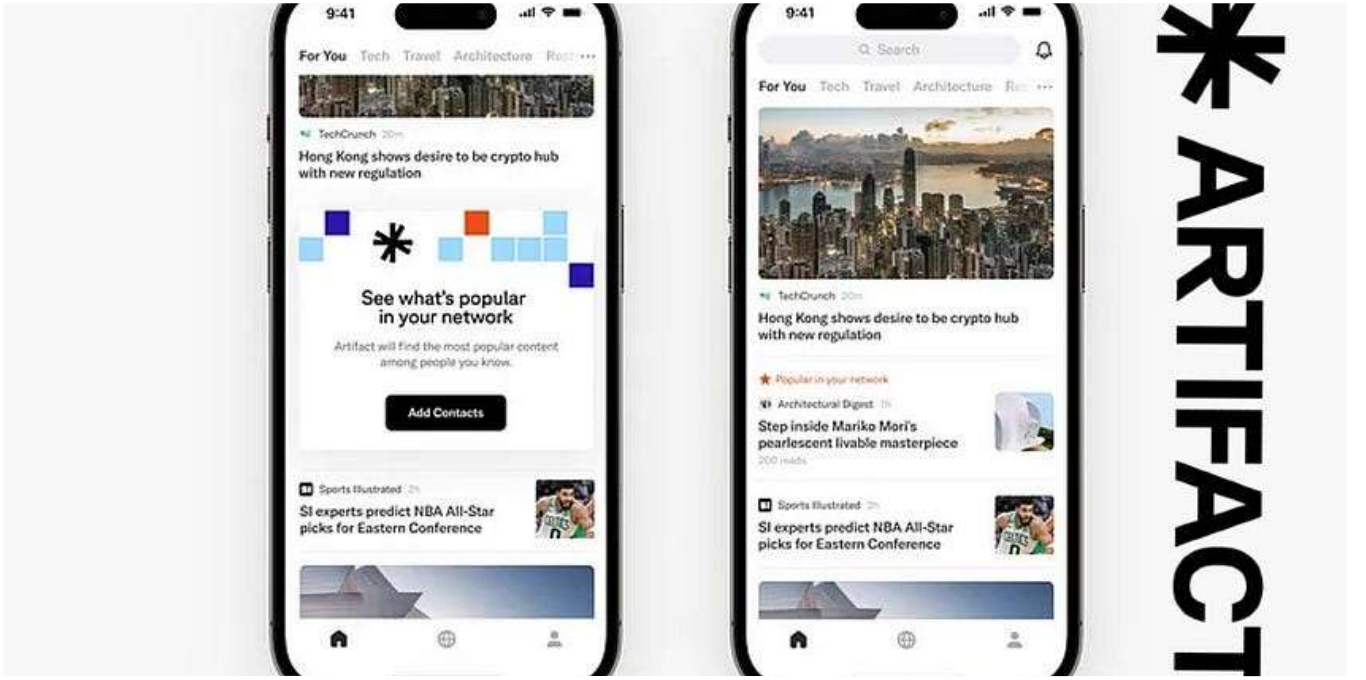
2




See all from Hakkı Kaan Simsek

See all from Veri Bilimi Türkiye

Recommended from Medium



 Gowtham Oleti

Apps I Use And Why You Should Too.

Let's skip past the usual suspects like YouTube, WhatsApp and Instagram. I want to share with you some less familiar apps that have become...

10 min read · Nov 14, 2023

 11.8K  207

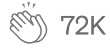


 Unbecoming

10 Seconds That Ended My 20 Year Marriage

It's August in Northern Virginia, hot and humid. I still haven't showered from my morning trail run. I'm wearing my stay-at-home mom...

★ • 4 min read • Feb 16, 2022



72K



1033



Lists



Staff Picks

548 stories • 599 saves



Stories to Help You Level-Up at Work

19 stories • 397 saves



Self-Improvement 101

20 stories • 1146 saves



Productivity 101

20 stories • 1047 saves



Sheila Teo in Towards Data Science

How I Won Singapore's GPT-4 Prompt Engineering Competition

A deep dive into the strategies I learned for harnessing the power of Large Language Models

★ • 24 min read • 6 days ago



2.9K



41



Alexandru Lazar in ILLUMINATION

Ten Habits that will get you ahead of 99% of People

Improve your life and get ahead of your peers in 10 simple steps

9 min read • Nov 18, 2023



15.8K



281





Michelle A. Cmarik in Age of Empathy

The Last Night I Slept With My Husband

Sleeping side-by-side had become mundane, until it felt momentous

★ · 5 min read · Dec 21, 2023



7.2K



109



Pragmatic Coders

AI predictions: Top 13 AI trends for 2024

Explore the future with our comprehensive guide to the top 13 AI trends anticipated for 2024.

13 min read · Dec 10, 2023

 4.6K  119

See more recommendations