Makine Öğrenmesi Dersleri 5a: Random Forest (Sınıflandırma)

Hakkı Kaan Simsek · Follow

Published in Veri Bilimi Türkiye

4 min read · Mar 24, 2018

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 başarılı bir karar ağacı günlük hayatta bilgi kazancını arttıracak 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 x sayıda öznitelik verip en faydalı y 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: 32561 entries, 0 to 32560
Data columns (total 15 columns):
age               32561 non-null int64
workclass         32561 non-null object
fnlwgt            32561 non-null int64
education         32561 non-null object
education-num     32561 non-null int64
marital-status    32561 non-null object
occupation        32561 non-null object
relationship      32561 non-null object
race              32561 non-null object
sex               32561 non-null object
capital-gain      32561 non-null int64
capital-loss      32561 non-null int64
hours-per-week    32561 non-null int64
native-country    32561 non-null object
salary            32561 non-null object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB
None
```

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	salary
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

İlk derste olduğu gibi sütunları isimlendirip veri tiplerini ve eksik veri olup olmadığını kontrol ettik. Modelimizdeki hedef değişken ‘salary’ olduğu için onun dağılımına bakıyoruz ve veri setinin geri kalanından ayırıyoruz.

```
Salary Distribution:
 <=50K    24720
 >50K      7841
Name: salary, dtype: int64
```

Bu çalışmada işimize yaramasa da yanlış bir veri tipi olduğunda nasıl düzeltmemiz gerektiğine bir örnek yaptım. Capital-gain değişkeninin tipini ‘integer’ dan ‘float’ a çevirdim.

Şimdi de tipi object olan verilere hızlıca bir göz atalım.

	workclass	education	marital-status	occupation	relationship	race	sex	native-country
32541	?	HS-grad	Separated	?	Not-in-family	Black	Female	United-States
32542	?	HS-grad	Married-civ-spouse	?	Husband	White	Male	United-States
32543	Local-gov	Assoc-acdm	Divorced	Prof-specialty	Unmarried	White	Female	United-States
32544	Private	Masters	Divorced	Other-service	Not-in-family	Other	Female	United-States
32545	Local-gov	Assoc-acdm	Married-civ-spouse	Adm-clerical	Wife	White	Female	United-States
32546	Private	Assoc-acdm	Divorced	Tech-support	Not-in-family	White	Female	United-States
32547	Private	HS-grad	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	Mexico
32548	Self-emp-not-inc	Prof-school	Never-married	Prof-specialty	Not-in-family	White	Male	United-States
32549	State-gov	Some-college	Divorced	Adm-clerical	Other-relative	White	Female	United-States
32550	Self-emp-not-inc	Some-college	Married-civ-spouse	Craft-repair	Husband	White	Male	United-States
32551	Private	10th	Married-civ-spouse	Handlers-cleaners	Husband	Amer-Indian-Eskimo	Male	United-States
32552	Private	Assoc-voc	Married-civ-spouse	Sales	Husband	White	Male	United-States
32553	Private	Masters	Never-married	Tech-support	Not-in-family	Asian-Pac-Islander	Male	Taiwan
32554	Private	Masters	Married-civ-spouse	Exec-managerial	Husband	White	Male	United-States
32555	Private	Some-college	Never-married	Protective-serv	Not-in-family	White	Male	United-States
32556	Private	Assoc-acdm	Married-civ-spouse	Tech-support	Wife	White	Female	United-States
32557	Private	HS-grad	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	United-States
32558	Private	HS-grad	Widowed	Adm-clerical	Unmarried	White	Female	United-States
32559	Private	HS-grad	Never-married	Adm-clerical	Own-child	White	Male	United-States
32560	Self-emp-inc	HS-grad	Married-civ-spouse	Exec-managerial	Wife	White	Female	United-States

Gördüğünüz gibi bazı veriler eksik olarak gözükmese de ‘?’ soru işareti ile doldurulmuş.

Yapmamız gerekenler sırasıyla şöyle:

- Veri tipi ‘object’ olan sütunlar seçilir.
- Bu sütunlar yine o sütunun mode() değeriyle doldurulur.

	workclass	education	marital-status	occupation	relationship	race	sex	native-country
32541	Private	HS-grad	Separated	Prof-specialty	Not-in-family	Black	Female	United-States
32542	Private	HS-grad	Married-civ-spouse	Prof-specialty	Husband	White	Male	United-States
32543	Local-gov	Assoc-acdm	Divorced	Prof-specialty	Unmarried	White	Female	United-States
32544	Private	Masters	Divorced	Other-service	Not-in-family	Other	Female	United-States
32545	Local-gov	Assoc-acdm	Married-civ-spouse	Adm-clerical	Wife	White	Female	United-States
32546	Private	Assoc-acdm	Divorced	Tech-support	Not-in-family	White	Female	United-States
32547	Private	HS-grad	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	Mexico
32548	Self-emp-not-inc	Prof-school	Never-married	Prof-specialty	Not-in-family	White	Male	United-States
32549	State-gov	Some-college	Divorced	Adm-clerical	Other-relative	White	Female	United-States
32550	Self-emp-not-inc	Some-college	Married-civ-spouse	Craft-repair	Husband	White	Male	United-States
32551	Private	10th	Married-civ-spouse	Handlers-cleaners	Husband	Amer-Indian-Eskimo	Male	United-States
32552	Private	Assoc-voc	Married-civ-spouse	Sales	Husband	White	Male	United-States
32553	Private	Masters	Never-married	Tech-support	Not-in-family	Asian-Pac-Islander	Male	Taiwan
32554	Private	Masters	Married-civ-spouse	Exec-managerial	Husband	White	Male	United-States
32555	Private	Some-college	Never-married	Protective-serv	Not-in-family	White	Male	United-States
32556	Private	Assoc-acdm	Married-civ-spouse	Tech-support	Wife	White	Female	United-States
32557	Private	HS-grad	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	United-States
32558	Private	HS-grad	Widowed	Adm-clerical	Unmarried	White	Female	United-States
32559	Private	HS-grad	Never-married	Adm-clerical	Own-child	White	Male	United-States
32560	Self-emp-inc	HS-grad	Married-civ-spouse	Exec-managerial	Wife	White	Female	United-States

Makine öğrenmesi modelleri kategorik değişkenleri algılayamadığı için ‘object’ tipindeki değişkenleri one-hot-encoding yöntemiyle 0 ve 1'lere ayırıyoruz.

	age	fnlwgt	education-num	capital-gain	capital-loss	hours-per-week	workclass_ Federal-gov	workclass_ Local-gov	workclass_ Never-worked	workclass_ Private	...	native-country_ Portugal	native-country_ Puerto-Rico	native-country_ Scotland	native-country_ South	native-country_ Taiwan
0	39	77516	13	2174	0	40	0	0	0	0	...	0	0	0	0	0
1	50	83311	13	0	0	13	0	0	0	0	...	0	0	0	0	0
2	38	215646	9	0	0	40	0	0	0	1	...	0	0	0	0	0
3	53	234721	7	0	0	40	0	0	0	1	...	0	0	0	0	0
4	28	338409	13	0	0	40	0	0	0	1	...	0	0	0	0	0

5 rows × 105 columns

Modelimizi kurmaya hazırız şimdi standart modelleme süreçlerini uygulayacağız.

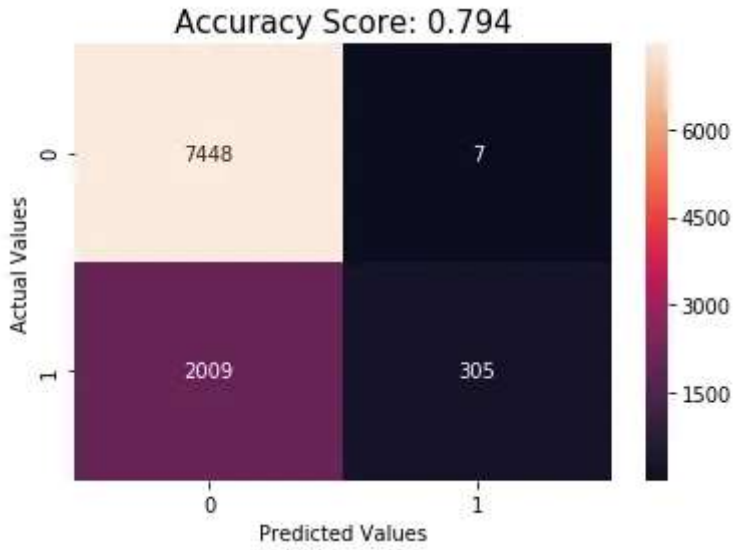
1. Veriyi eğitim ve test alt-veri gruplarına ayırma.
2. Karar ağacı modeli oluşturma.

3. Modeli eğitim verisine ‘fit’ etme.

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            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.

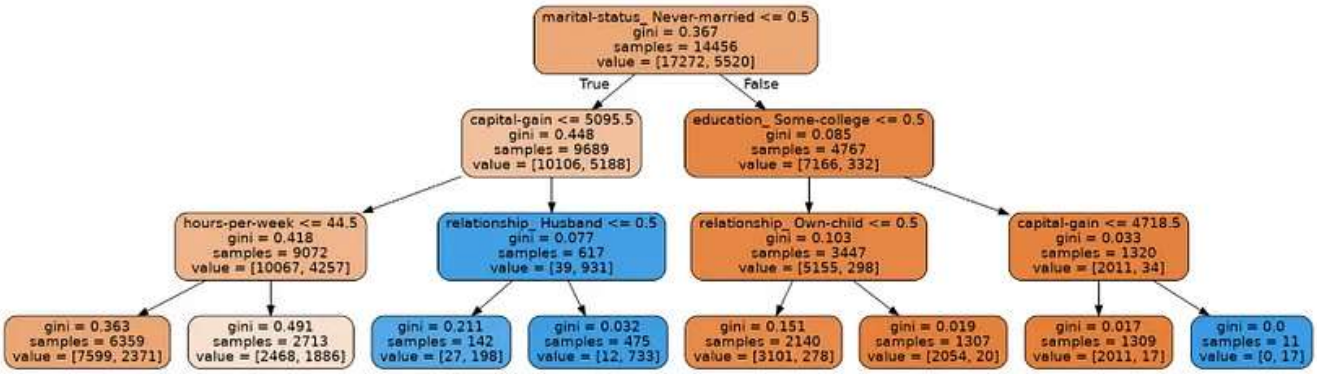
5a. Modelin başarı metrikleri: Confusion matrix


Accuracy Score: 0.794
7448
7
2009
305
0
1
Actual Values
0
1
Predicted Values
6000
4500
3000
1500

5b. Modelin başarı metrikleri: Precision, recall, f1-score

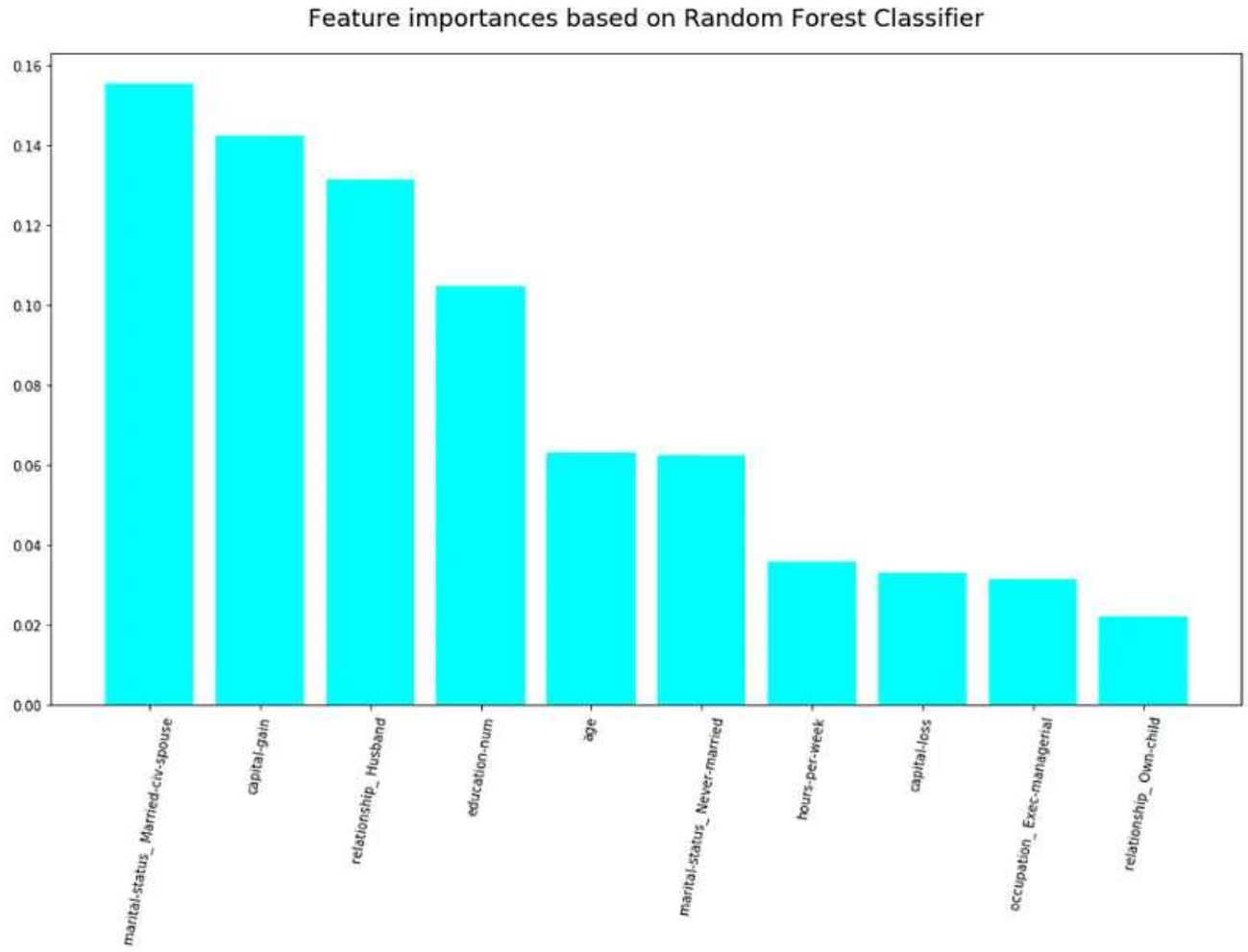
	precision	recall	f1-score	support
<=50K	0.79	1.00	0.88	7455
>50K	0.98	0.13	0.23	2314
avg / total	0.83	0.79	0.73	9769

6. Karar ağaçlarından birini görselleştirme.



7. Modelin özniteliklerinin önem sıralamasını analiz etme.

	precision	recall	f1-score	support
<=50K	0.87	0.96	0.91	7455
>50K	0.79	0.53	0.63	2314
avg / total	0.85	0.86	0.84	9769



Bu çalışma standart bir veri bilimi projesinin en basit hali olarak düşünülebilir.

Veri tiplerini kontrol etme/düzeltme.

- 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.
- Modelin yaptığı öznitelik sıralamasını görselleştirme.

Çalışmadaki veri setine ve kodlara şuradan ulaşabilirsiniz.

Sorunuz olursa bana Linkedin veya Twitter hesaplarından yazabilirsiniz.

Buyuk Veri | Büyük Veri | Makine Öğrenmesi | Yapay Zeka | Yapay Öğrenme

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 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

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 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

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:

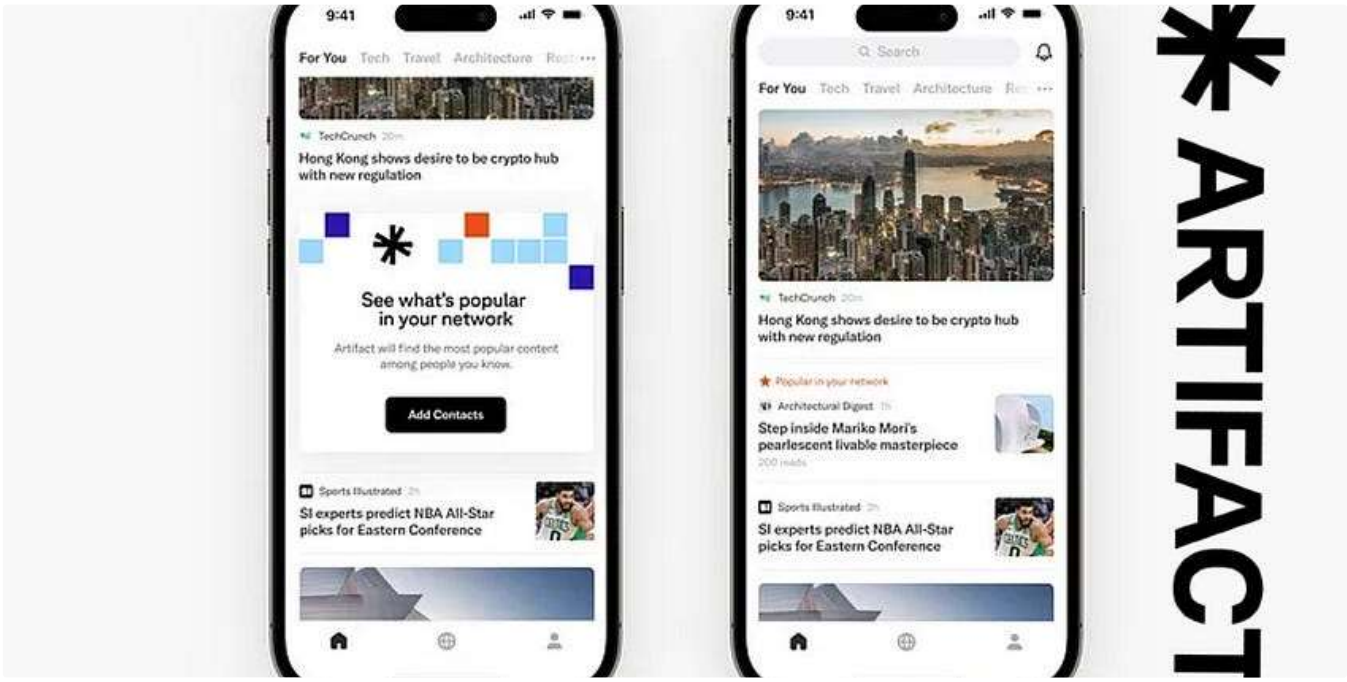
4 min read · Mar 6, 2018

303 3

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 · 1046 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

2.8K 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