# Makine Öğrenmesi Dersleri 9: Hiper-parametre Kestirimi

Hakkı Kaan Simsek · Follow
Published in Veri Bilimi Türkiye
3 min read · May 6, 2018



kaynak

Hiper-parametre seçimi (hyper-parameter tuning) elinizle radyo frekansı ayarlamaya benziyor. Hani ses iyidir ama siz bi tık daha iyi olmasını istersiniz ya işte hiper-parametre seçimi de makine öğrenmesi modelleri için o işe yarıyor.

Örnek veri seti olarak elimizde Portekiz bankasından alınmış 11 bin kişinin yaşı, mesleği, medeni durumu, ev kredisi alıp almadığı, son görüşmeden sonra geçen zaman, görüşmenin sabit telefonla mı cep telefonuyla mı gerçekleştiği gibi öznitelikler var. Kişilerin bankaya depozito yatırıp yatırmayacağını tahmin eden bir model kurmaya çalışıyoruz. Çıkacak sonuçlara göre belirli kişilere ve gruplara kişiselleştirilmiş pazarlama yöntemleri uygulanabilir.

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('https://raw.githubusercontent.com/HakkiKaanSimsek/Makine_Ogrenmesi_Dersleri/ma
print(df.info())
print('')
df.head(10)
```

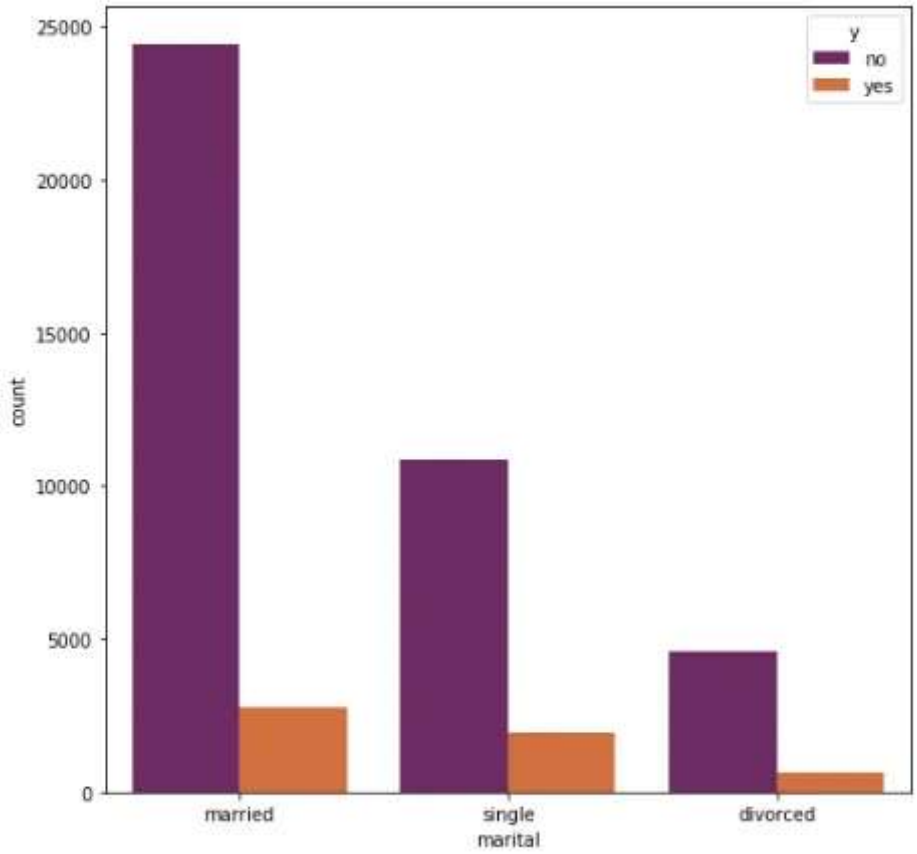
Makine Öğrenmesi Dersleri-8.py hosted with ❤ by GitHub view raw

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
age          45211 non-null int64
job          45211 non-null object
marital      45211 non-null object
education    45211 non-null object
default      45211 non-null object
balance      45211 non-null int64
housing      45211 non-null object
loan         45211 non-null object
contact      45211 non-null object
day          45211 non-null int64
month        45211 non-null object
duration     45211 non-null int64
campaign     45211 non-null int64
pdays        45211 non-null int64
previous     45211 non-null int64
poutcome     45211 non-null object
y            45211 non-null object
dtypes: int64(7), object(10)
memory usage: 5.9+ MB
None
```

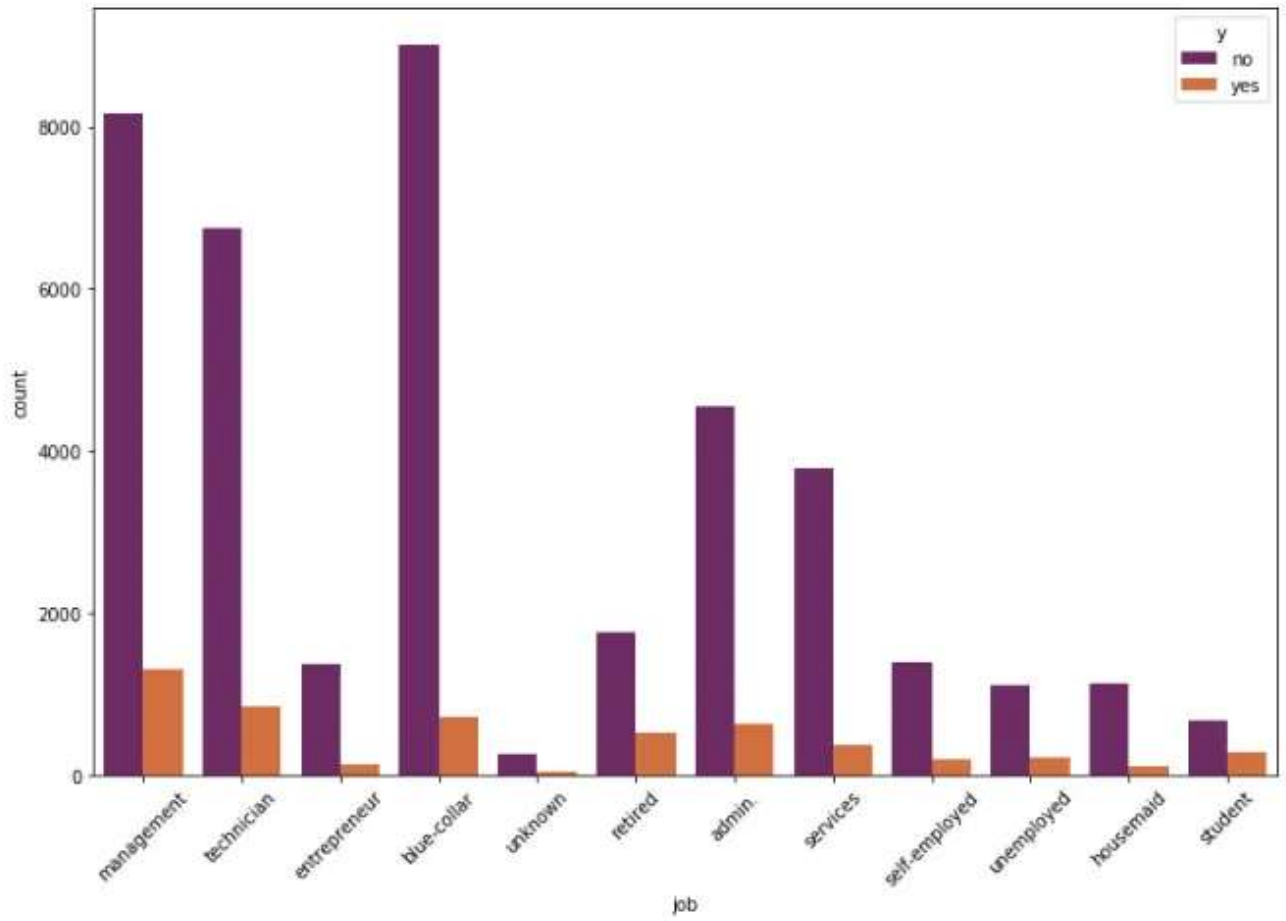
| | age | job | marital | education | default | balance | housing | loan | contact | day | month | duration | campaign | pdays | previous | poutcome | y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 58 | management | married | tertiary | no | 2143 | yes | no | unknown | 5 | may | 261 | 1 | -1 | 0 | unknown | no |
| 1 | 44 | technician | single | secondary | no | 29 | yes | no | unknown | 5 | may | 151 | 1 | -1 | 0 | unknown | no |
| 2 | 33 | entrepreneur | married | secondary | no | 2 | yes | yes | unknown | 5 | may | 76 | 1 | -1 | 0 | unknown | no |
| 3 | 47 | blue-collar | married | unknown | no | 1506 | yes | no | unknown | 5 | may | 92 | 1 | -1 | 0 | unknown | no |
| 4 | 33 | unknown | single | unknown | no | 1 | no | no | unknown | 5 | may | 198 | 1 | -1 | 0 | unknown | no |
| 5 | 35 | management | married | tertiary | no | 231 | yes | no | unknown | 5 | may | 139 | 1 | -1 | 0 | unknown | no |
| 6 | 28 | management | single | tertiary | no | 447 | yes | yes | unknown | 5 | may | 217 | 1 | -1 | 0 | unknown | no |
| 7 | 42 | entrepreneur | divorced | tertiary | yes | 2 | yes | no | unknown | 5 | may | 380 | 1 | -1 | 0 | unknown | no |
| 8 | 58 | retired | married | primary | no | 121 | yes | no | unknown | 5 | may | 50 | 1 | -1 | 0 | unknown | no |
| 9 | 43 | technician | single | secondary | no | 593 | yes | no | unknown | 5 | may | 55 | 1 | -1 | 0 | unknown | no |

```python
plt.figure(figsize=(8,8))
sns.countplot(x = 'marital', hue = 'y', data=df,palette = 'inferno')
plt.show()
```

Makine Öğrenmesi Dersleri-8.py hosted with ❤ by GitHub view raw




Bekar insanların evli insanlara kıyasla bankada vadeli para tutma oranının fazla olduğunu görüyoruz.



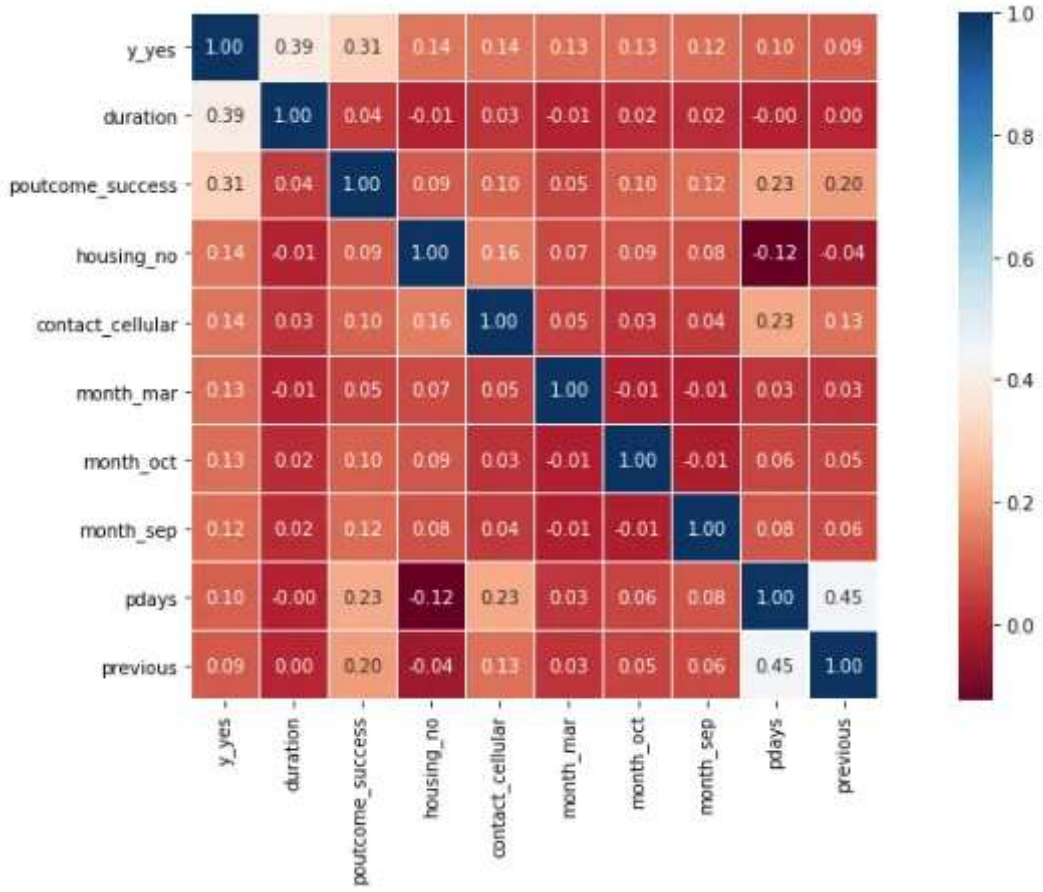

Yukarıdaki tabloda meslek dağılımlarına göre kişilerin bankaya vadeli para yatırıp yatırmadıklarını görüyoruz. Örneğin mavi yakalılarda durum hiç açıcı değil. Buradan hareketle farklı pazarlama kampanyaları düşünülebilir.

Son olarak depozito yatırmakla diğer değişkenler arasındaki korelasyona bakmak istiyoruz. **One-hot-encoder yöntemiyle kategorik değişkenleri binary (0,1) hale getiriyoruz aksi halde model ne cinsiyeti ne medeni durumu ne de iş güç hiçbir şey anlamaz.**

| | age | balance | day | duration | campaign | pdays | previous | job_admin. | job_blue-collar | job_entrepreneur | ... | month_mar | month_may | month_nov | month_oct |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 58 | 2143 | 5 | 261 | 1 | -1 | 0 | 0 | 0 | 0 | ... | 0 | 1 | 0 | 0 |
| 1 | 44 | 29 | 5 | 151 | 1 | -1 | 0 | 0 | 0 | 0 | ... | 0 | 1 | 0 | 0 |
| 2 | 33 | 2 | 5 | 76 | 1 | -1 | 0 | 0 | 0 | 1 | ... | 0 | 1 | 0 | 0 |
| 3 | 47 | 1506 | 5 | 92 | 1 | -1 | 0 | 0 | 1 | 0 | ... | 0 | 1 | 0 | 0 |
| 4 | 33 | 1 | 5 | 198 | 1 | -1 | 0 | 0 | 0 | 0 | ... | 0 | 1 | 0 | 0 |

5 rows × 52 columns

Gördüğünüz gibi satır sayısı 17'den 52'ye çıktı. 52*52 bir korelasyon matrisine bakmak pek mümkün değil onun için depozito yatırmakla en yüksek korelasyonu olan 10 değişkeni seçiyoruz.



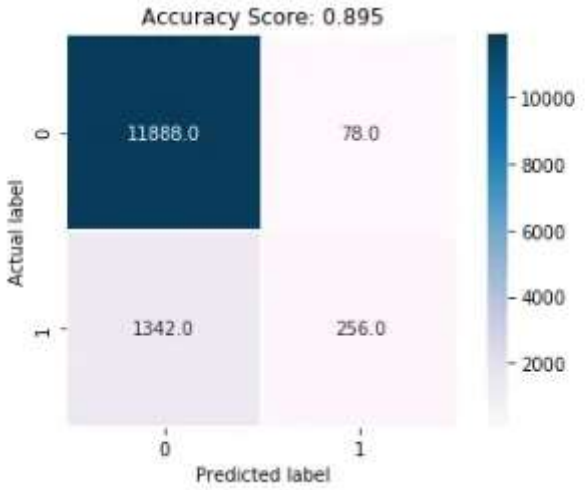

Depozito yatırmakla en yüksek pozitif korelasyonu olan 5 öznitelik:

- Son telefon konuşmasının uzunluğu (son konuşmanın süresi 0'sa kişi telefonda ikna edilmemiştir, o yüzden çok iyi bir öznitelik olmayabilir.)
- Bir önceki pazarlama kampanyasının başarılı sonuç vermesi.
- Ev kredisi alınmaması.
- Konuşmanın sabit telefon yerine cep telefonu ile gerçekleşmesi.

Bu yazının konusu hiper-parametre seçimi olduğu için açıklayıcı veri analizi kısmını burada bitiriyoruz ama gördüğünüz gibi sadece veriyi analiz ederek bile iyi bilgiler elde edebiliyoruz.

Hedef değişkenimizi tutuyoruz. Sonrasında veri setimizi train ve test setlerine ayırıyoruz, buraya kadar her şey oldukça basit şimdi random forest modelimizi kuralım.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| <=50K | 0.90 | 0.99 | 0.94 | 11966 |
| >50K | 0.77 | 0.16 | 0.27 | 1598 |
| avg / total | 0.88 | 0.90 | 0.86 | 13564 |




Sonuçlar kötü değil. Hiç görmediğimiz bir veri setinde %89.5 olasılıkla doğru tahminler yapıyoruz. Geçen hafta öğrendiğimiz cross-validation metodunu uygulayalım şimdi de.

```
[0.89717264 0.89650814 0.89666614 0.89508611 0.89398009]

mean of cv-scores 0.8961
```

Modelin 5 farklı eğitim ve test veri setindeki ortalaması %89.6 yani az önce şans eseri %89.5 gibi bir doğruluk oranı yakalamadık modelimiz gayet iyi çalışıyor.

Daha gidecek yol var mı bi bakalım...

```
Fitting 5 folds for each of 100 candidates, totalling 500 fits
[Parallel(n_jobs=-1)]: Done 120 tasks      | elapsed:   10.7s
[Parallel(n_jobs=-1)]: Done 370 tasks      | elapsed:   40.1s
[Parallel(n_jobs=-1)]: Done 500 out of 500 | elapsed:   56.3s finished

{'max_depth': 12, 'max_features': 14}
Best cv mean result: 0.90641
Best holdout result: 0.90467
```

GridSearch bizim için 500 farklı kombinasyonu deneyerek en iyi sonuçlar veren hiper-parametreleri seçti. **(max_depth:12, max_features:14)**

Bu hiper-parametre kombinasyonları denenerek bir model kurulduğunda en iyi doğruluk oranı %90.6, bizim ilk modeli kurduğumuz eğitim ve test veri setindeki

doğruluk oranı %90.4. **(ilk model: %89.5)**

Yukarıdaki adımlar izlenerek diğer makine öğrenmesi modellerinde de basitçe hiper-parametre kestirimi yapılabilir.

Örneğin knn için **params = {'knn__n_neighbors': range(1, 5)}** yazarak **GridSearchCV** yapabilirsiniz.

Çalışmadaki veri setine ve kodlara şuradan ulaşabilirsiniz.

Sorunuz olursa bana Linkedin veya Twitter hesaplarından yazabilirsiniz.

Buyuk Veri | Büyük Veri | Makine Öğrenmesi | Derin Ogrenme | Yapay Zeka

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 5a: Random Forest (Sınıflandırma)

Rassal orman (Random Forest), hiper parametre kestirimi yapılmadan da iyi sonuçlar vermesi hem regresyon hem de sınıflandırma problemlerine...

4 min read · Mar 24, 2018

390 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 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

See all from Hakkı Kaan Simsek

See all from Veri Bilimi Türkiye

## Recommended from Medium

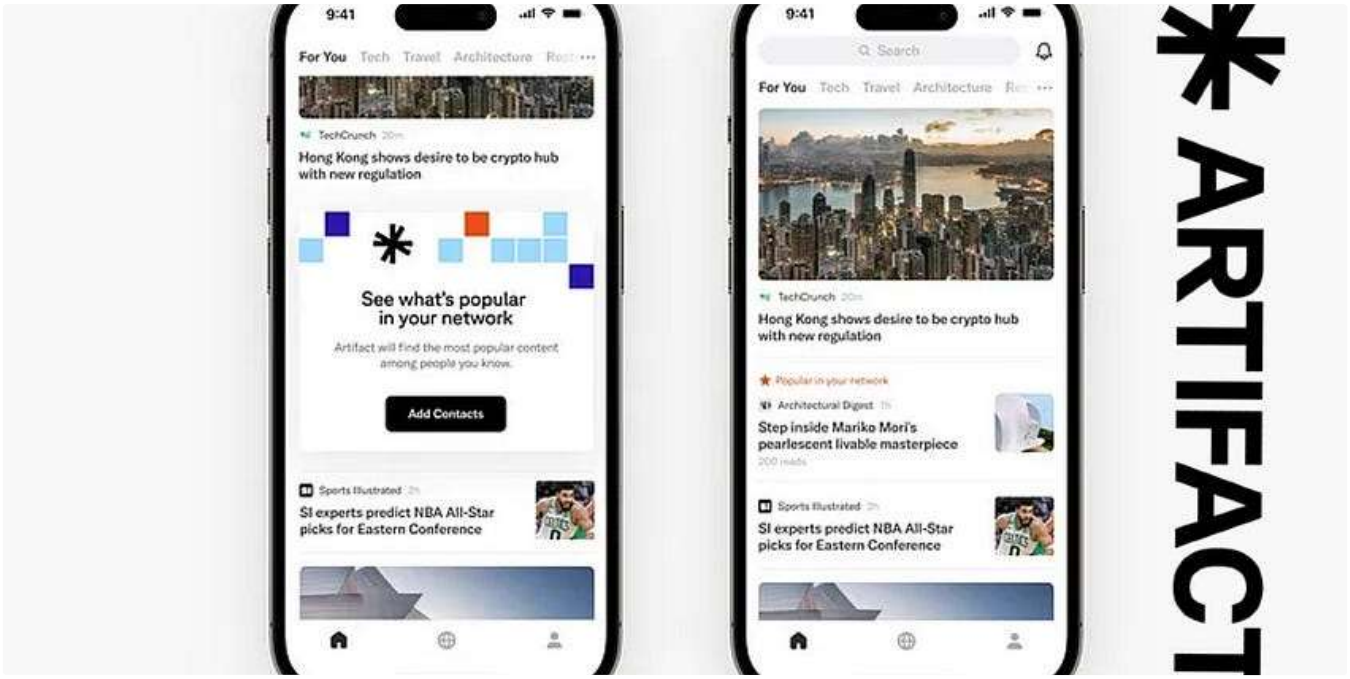
Unbecoming

## 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

71K 1026




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

10.3K 184

### Lists

**Staff Picks**

543 stories · 577 saves

**Stories to Help You Level-Up at Work**

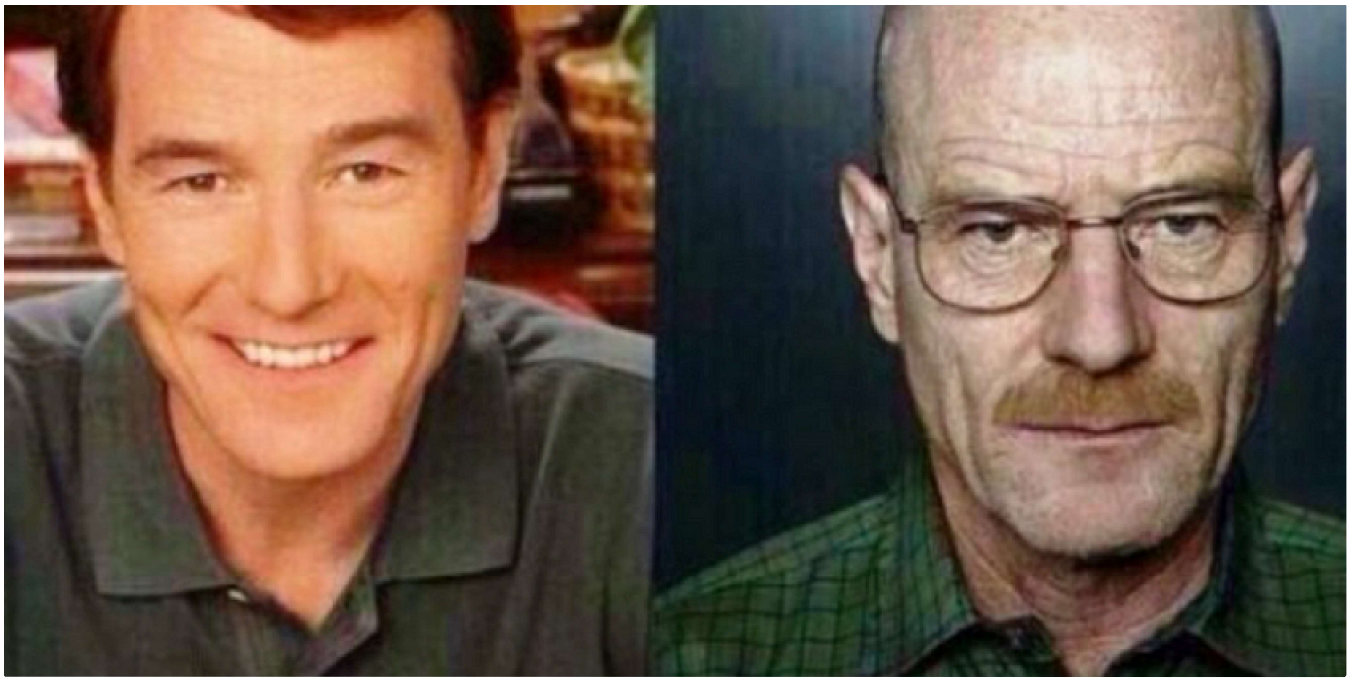
19 stories · 385 saves

**Self-Improvement 101**

20 stories · 1104 saves

**Productivity 101**

20 stories · 1012 saves



David Goudet

## This is Why I Didn't Accept You as a Senior Software Engineer

An Alarming Trend in The Software Industry

· 5 min read · Jul 26

7.2K 75

Scott-Ryan Abt in Pitfall

## Bye Bye, Spotify

And see ya later, all you subscription services in my little empire

· 4 min read · Aug 19

17.9K 425

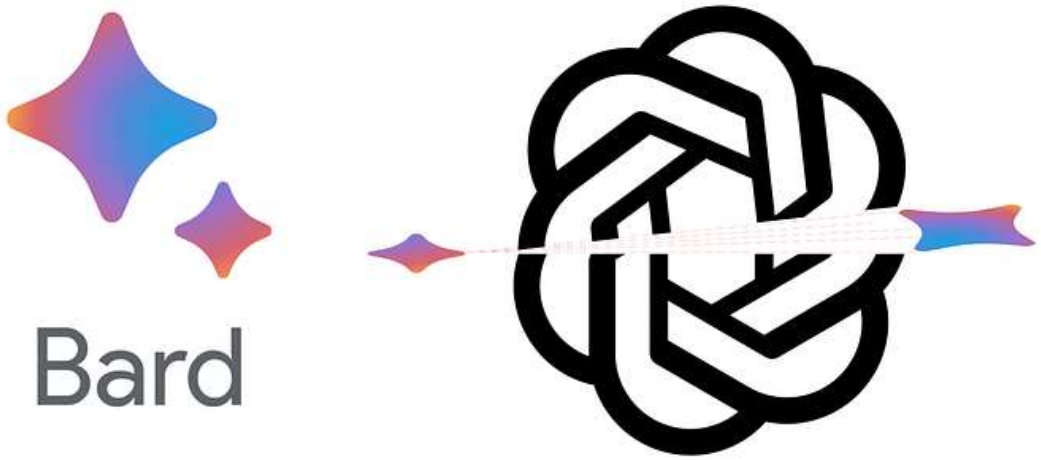
Alexandru Lazar in ILLUMINATION

## 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

14.2K 259




AL Anany

## The ChatGPT Hype Is Over—Now Watch How Google Will Kill ChatGPT.

It never happens instantly. The business game is longer than you know.

✦ · 6 min read · Sep 1

20K 637

See more recommendations