🙄

Hafta 6 (Feature Engineering)

@mebaysan

01/10/2021

Bu haftanın veri seti

-

İlgili Okuma Listesi

-

Benim Yazdığım Yazılar

- What is Feature Engineering?

Feature Engineering & Data Pre-Processing (Özellik Mühendisliği & Veri Ön İşleme)

Verimiz kötü ise modellerimizin çıktısı kötü olacaktır.

If Your Data Is Bad, Your Machine Learning Tools Are Useless

by Thomas C. Redman

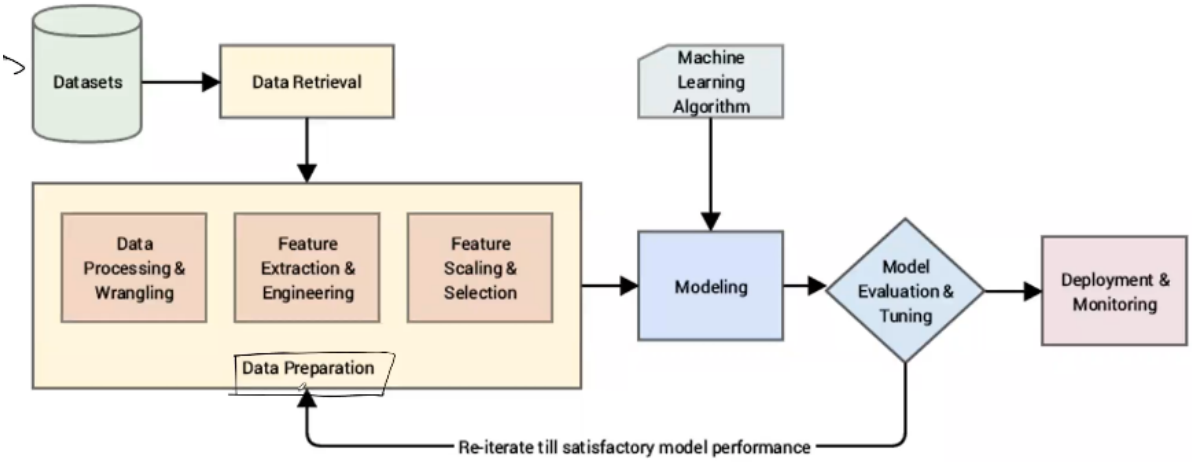
April 02, 2018

Alan Schein Photography/Getty Images

> Applied machine learning is basically feature engineering
> - Andrew NG

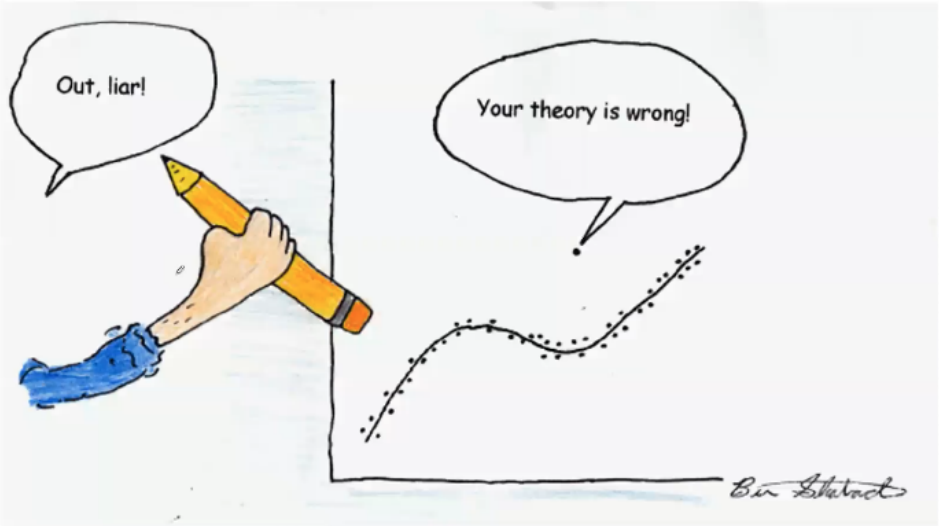
Özellik Mühendisliği: Özellikler üzerinde gerçekleştirilen çalışmalardır. Ham veriden değişken üretmektir.

Veri Ön İşleme: Çalışmalar öncesi verinin uygun hale getirilmesidir.



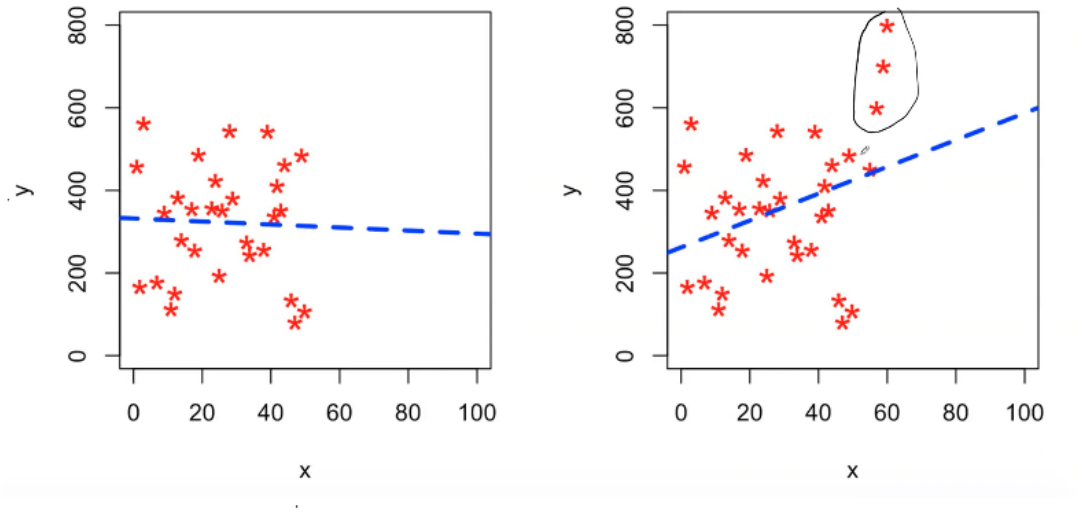
csaybar.githu

Outliers (Aykırı Değerler)



Verideki genel eğilimin oldukça dışına çıkan değerlere aykırı değer denir.

Aşağıdaki örnekte gördüğümüz gibi, aykırı değerler (sağda) değişkenler arasındaki ilişkileri bozabilir.



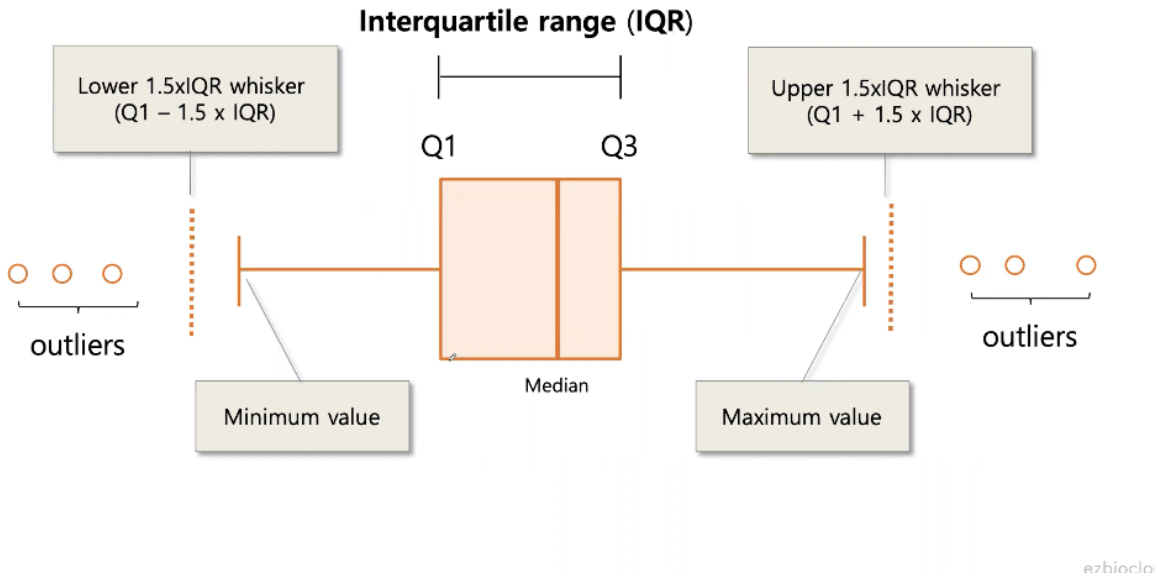
Doğrusal yöntemler aykırı değerlerden daha çok etkilenir.

Ağaç yöntemleri aykırı değerlerden daha az etkilenir.

Aykırı değerler bu şekillerde ölçülebilir:

- Sektör bilgisi
  - Örnek: ev fiyat tahmin modeli geliştirirken biliriz ki 1000 m2 görülme frekansı düşüktür, bu aykırı değerdir diyerek bunu veriden çıkarabilir veya çevirebiliriz
- Standart Sapma Yaklaşımı
  - Belirli bir standart sapma yukarıda olan değişkenlere aykırı değer muamelesi yapılabilir.
- Z-Skoru Yaklaşımı
  - Değişken standartlaştırılır, mesela -3 ile +3 arasında değişken standartlaştırılır ve limitlerin üzerinde olan değerlere aykırı değer muamelesi yapabiliriz.

- Boxplot (interquartile range - IQR) Yöntemi
  - Boxplot limitlerinin (Q1 - 1.5*IQR) ile (Q3 + 1.5*IQR) dışında kalan değerlere aykırı değer muamelesi yapabiliriz
  - IQR: Q3 - Q1
  - Vahit Hoca Q1 için 0.05 ve Q3 için 0.95 kullanıyor. Sadece uçlardan traşlamanın verinin yapısını bozmadığını düşünüyor.



Aykırı Değer Problemini bu şekillerde çözebiliriz:

- Silmek
  - Veri setinden aykırı değere sahip olan gözlemleri çıkarırız
  - Aykırı değerleri veri setinden çıkarmak yapıyı bozabileceğinden dolayı genelde bu yöntem pek fazla tercih edilmez.
- Baskılama Yöntemi (re-assignment with thresholds)

- UP limit üstünde olan değerlere UP limit değeri atanır
- LOW limit altında olan değerlere LOW limit değeri atanır

Çok Değişkenli Aykırı Değer Analizi: Local Outlier Factor

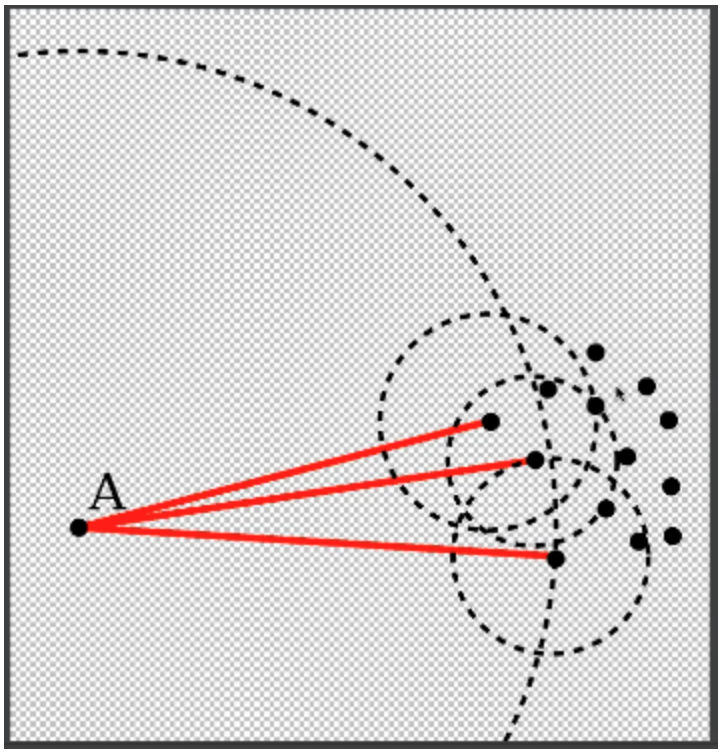
"Yaş" ve "Hamilelik Sayısı" adında 2 değişkenimiz olduğunu varsayalım.

- Tek başına "17" değeri "Yaş" değişkeni için aykırı bir değer değildir
- Tek başına "3" değeri "Hamilelik Sayısı" değişkeni için aykırı bir değer değildir
- Fakat "17 yaşında 3 kere hamilelik sayısı" olma gözlemi bir aykırı değerdir.

Çok değişkenli aykırı değer analizi bize; değişkenlerin eş anlı etkilerini değerlendirme imkanı verir. Yani normalde tek başına aykırı değer olamayacak değerler bir araya geldiğinde aykırı değer oluşturuyor olabilir.

"Local Outlier Factor" çok değişkenli aykırı değer analizi yöntemidir.

Temelde yaptığı iş: gözlemleri bulundukları konumda yoğunluk tabanlı skorlayarak, aykırı değer olabilecek değerleri tanımlama imkanı sağlamasıdır.



Python'da hesaplamak için:

```
from sklearn.neighbors import LocalOutlierFactor
```

Missing Values (Eksik Değerler)

Gözlemlerde eksiklik olması durumunu ifade etmektedir.

Bu problemi şu şekillerde çözebiliriz:

- Silme
  - Eksik verisi olan gözlemleri uçurmayı ifade eder
- Değer Atama Yöntemi
  - Ortalama, mod, medyan gibi değerleri eksik değerlere atayabiliriz
- Tahmine Dayalı Yöntemler
  - Makine öğrenmesi (KNN) veya istatistiksel iterasyonlu yöntemler ile eksik gözlemleri doldurmayı ifade eder

> "The idea of imputation is both seductive and dangerous"
> - R.J.A Little & D.B Rubin

Eksik veri ile çalışılırken göz önünde bulundurulması gereken en önemli konulardan birisi eksik verinin rassallığıdır.

> Eksik değere sahip gözlemlerin veri setinden direk çıkarılması ve rassallığının incelenmemesi, yapılacak istatistiksel çıkarımların ve modelleme çalışmalarının güvenilirliğini düşürecektir.
> - Alpar, 2011

Eksik gözlemlerin veri setinden direk çıkarılabilmesi için veri setindeki eksikliğin bazı durumlarda *kısmen* bazı durumlarda *tamamen rastlantısal* olarak oluşmuş olması gerekmektedir.

Eğer eksiklikler değişkenler ile ilişkili olarak ortaya çıkan *yapısal problemler* ile meydana gelmiş ise bu durumda yapılacak silme işlemleri ciddi yanlılıklara sebep olabilecektir.

(Tabachnick ve Fidell, 1996)

Encoding (Label Encoding, One-Hot Encoding, Rare Encoding)

Değişkenlerin temsil şekillerinin değiştirilmesidir.

Label Encoding

Label encoding'i 1 ve 0 olarak yaparsak binary encoding yapmış oluruz.

Aşağıdaki örnekte "Kadın mı?" olarak encoding yaptık. Kadınsa 1 değilse 0 olacak şekilde encode ettik.

SEX	IS_FEMALE
Male	0
Female	1
Female	1
Male	0
Male	0
Female	1

Nominal: Sınıflar arası fark yok demektir.

Ordinal: Sınıflar arası fark var demektir.

Aşağıdaki örnek için EDUCATION değişkenini bu şekilde encode etmemiz doğrudur. Çünkü bu değişken ordinaldir.

EDUCATION	EDU_LABEL
Pre-School	0
Secondary School	1
High School	2
Graduate	3
Master	4
PhD	5

Fakat; aşağıdaki TEAM değişkeni nominal olduğundan dolayı bu şekilde encode etmemiz doğru değildir.

TEAM	EDU_LABEL
GS	0
FB	1
BJK	2
TS	3
BC	4
RM	5

One-Hot Encoding

Elimizdeki kategorik değişken nominal de olsa ordinal de olsa One-Hot Encoding kullanabiliriz.

Kategorik değişkenin sınıflarını bir değişkene dönüştürürüz.

TEAM	GS	FB	BJK	TS	BC	RM
GS	**1**	0	0	0	0	0
FB	0	**1**	0	0	0	0
BJK	0	0	**1**	0	0	0
TS	0	0	0	**1**	0	0
BC	0	0	0	0	**1**	0
RM	0	0	0	0	0	**1**

Dummy Değişken Tuzağı: Değişkenlerin birbirleri üzerinden oluşturulması tehlikesidir.

> Örnek olarak: "Cinsiyet" değişkenini düşünebiliriz. Bu değişkeni one-hot encoder'dan geçirirsek erkek ve kadın olarak 2 yeni değişkenimiz olacak. Bu değişkenler birbirini tamamlıyor olacaklardır ve birbirleri üzerinden oluşturuluyor olabilecektir. Bu durum dummy değişken tuzağıdır.

Genel kanaat; dummy değişken tuzağını göz önünde bulundurarak one-hot encoding kullanırken 1 sınıfı dışarıda bırakıp bunun önüne geçiyor olacağız. Yukarıdaki örnek için: GS encode edilirken dışarıda bırakılır ve diğer değişkenlerin hiç biri olmadığı durumda bu gözlemin GS olduğunu anlayabiliriz.

Rare Encoding

Bazı problemlerde elimizdeki kategorik değişkenlerin sınıf sayıları çok yüksek olabiliyor.

Aşağıdaki örnekte: Şehirler ve veri setindeki sıklıkları verilmiştir. Gördüğümüz üzere D-G-H-L gibi sınıfların

frekansları oldukça düşüktür. Bu kategorik değişkeni one-hot encoder'dan geçirerek sınıflarını bir değişken yapsak, bütün veri setinde sadece 1-3-5-2 defa geçen sınıflar (D-G-H-L) için değişken yapmaya gerek var mı? Yok. Rare encoder'daki temel amaç bu yorumdur. **Düşük frekanslı sınıflar için değişken oluşturmamaktır** diyebiliriz.

CITY	CITY_COUNT
A	56
B	84
C	54
D	2
E	12
F	60
G	3
H	5
K	25
L	1
M	36
Z	45

CITY	CITY_COUNT
A	56
B	84
C	54
D	2
E	12
F	60
G	3
H	5
K	25
L	1
M	36
Z	45

Bu örnek için senaryoya göre bir limit belirliyorum. 20'den az frekansı olan sınıfları rare olarak işaretliyorum ve aşağıdaki son halini alıyor. Bu şekliyle CITY değişkenini one-hot encoder'a soktuğumuzda gereksiz yere değişken oluşturmamış olacağız.

CITY	CITY_COUNT
A	56
B	84
C	54
D	2
E	12
F	60
G	3
H	5
K	25
L	1
M	36
Z	45

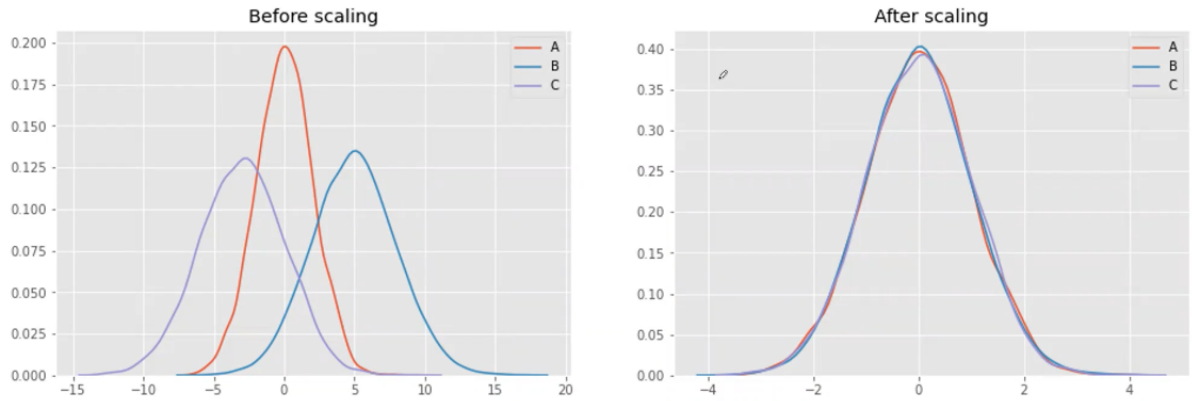
CITY	CITY_COUNT
A	56
B	84
C	54
D	2
E	12
F	60
G	3
H	5
K	25
L	1
M	36
Z	45

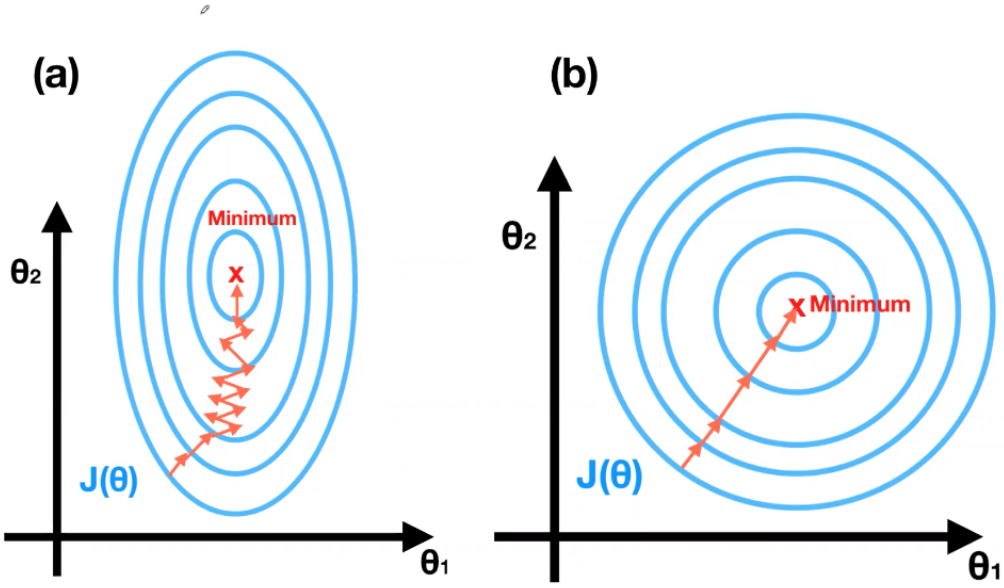
CITY	CITY_COUNT
A	56
B	84
C	54
F	60
K	25
M	36
Z	45
RARE (D,E,G,H,L)	23

Feature Scaling (Özellik Ölçeklendirme)

Değişkenler arasındaki ölçüm farklılığının önüne geçmek için yapılır.

Bazı yöntemlerde; ölçeklerin farklı olması değişkenlerin etkilerinin birbirlerine ezmesini sağlar. Aynı zamanda eğitim (train) süresini uzatır.





Standartlaştırma türleri:

- StandardScaler
  - Klasik standartlaştırma. Ortalamayı çıkar, standart sapmaya böl. z = (x - u) / s
- RobustScaler

    - Medyanı çıkar IQR'a böl.
- MinMaxScaler
    - Verilen 2 değer arasında değişken dönüşümü
- Log
    - Logaritmik dönüşüm.

Feature Extraction (Özellik Çıkarımı)

Ham veriden değişken üretmektir.

Örnek: Aşağıdaki veriden kadın erkek olup olmadığını çıkarabiliriz.