233. Kime Göre Neye Göre Aykırı Gözlem

Aykırı gözlemi veride genel eğilimin oldukça dışına çıkan gözlemler olarak tanımlamıştık. Peki veri setinin genel eğilimının dışına çıkmayı nasıl tanımlarız?

1. Sektör Bilgisi

Bazı senaryolarda aykırı gözlemleri nümerik yöntemlerle tanımlayamayız da çalışıyor olduğumuz sektörün dinamiklerine göre neyin aykırı ya da neyin normal olduğu durumunu tanımlayabiliriz.

Örneğin; bir ev fiyat tahmin modelinde 1000 metrekarelik evleri modellemeye almamak. Çünkü bizim makine öğrenmesi çalışmalarından amacımız genellenebilir, veri setinin içerisindeki yapıları temsil etme kabiliyeti yüksek, yansız modeller oluşturmak. Türkiye için depolandığında belki evlerin %0.99'unun 2+1, 3+1 veya daha küçük olduğunu biliyoruz. Dolayısıyla belki %1 bile olmayan 1000 metrekarelik evleri bir tahmin modeli için çalışmaya dahil etmek, diğer daha sık metrekarelerde yer alan evler için yapılacak olan genellemelere zarar verecektir.

Çünkü kullanıyor olduğumuz fonksiyonel yapılar oradaki 1000 metrekarelik evlerle ilişkili yapılarla alakalı da bazı genellenebilirlikler sağlamaya çalıştığından dolayı bu durumda veri setinin içerisindeki diğer yapıları kaçırabilmektedir.

2. Standart Sapma Yaklaşımı

Bir değişkenin ortalamasının üzerine aynı değişkenin standart sapması hesaplanarak eklenir. 1, 2 ya da 3 standart sapma değeri ortalama üzerine eklenerek ortaya çıkan bu değer eşik değer olarak düşünülür ve bu değerden yukarıda ya da aşağıda olan değerler aykırı değer olarak tanımlanır.

Diyelim ki elimizde araç fiyatları var. Bunların ortalaması ve standart sapma bilgisi var elimizde. Bunları kullanarak eşik değeri şu şekilde belirleyebiliriz.

$$\text{Eşik Değer} = \text{Ortalama} + 1 \times \text{Standart Sapma}$$
$$\text{Eşik Değer} = \text{Ortalama} + 2 \times \text{Standart Sapma}$$
$$\text{Eşik Değer} = \text{Ortalama} + 3 \times \text{Standart Sapma}$$

Hangisini kullanacağımız problemden probleme değişebilir.

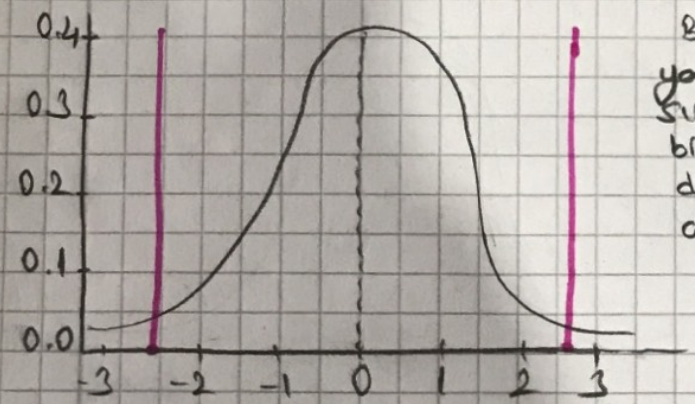
Örn; araç fiyat ortalaması 100.000 ve standart sapma = 20.000 ise bu durumda şunu diyebiliriz. Benim verimin merkezi eğilimini ben buldum 100.000. Üzerine bu değişkenin dağılımını (ortalama etrafındaki yapılanmasını) ifade eden standart sapmadan 2 tane ekledim.

$$\text{Eşik Değer} = 100{,}000 + 2 \times 20{,}000$$
$$= 140{,}000$$

3. Z-Skoru Yaklaşımı

Eğer elimizdeki değişken standartlaştırılıp buna göre bir muamele yapılacak bu durumda Z-skoru kullanılır.

Standart sapma yöntemine benzer şekilde çalışır. Değişken, standart normal dağılıma uyarlanır, yani standartlaştırılır. Sonrasından -örneğin- dağılımın sağından ve solundan -+2.5 değerine göre bir eşik değer konulur ve bu değerin üzerinde ya da altında olan değerler aykırı değer olarak işaretlenir.

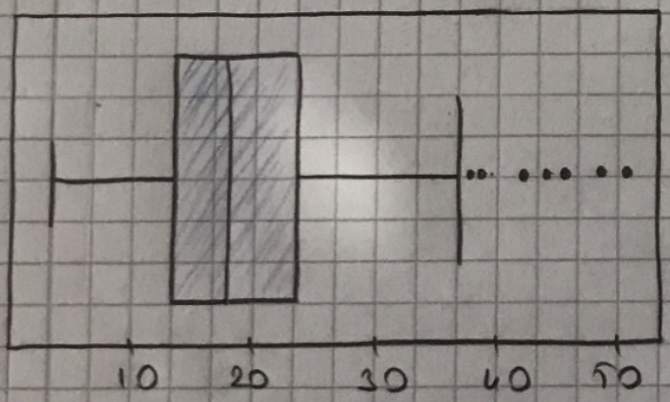


∓2.5

Buradaki yaklaşımda, değişkenin genel yapısının üzerine, değişkenin genel yapısının dağılımını da ifade eden +,- bir değer koyarak veri setinin genel dağılımının dışındaki yapılara erişilmeye çalışılır.

4. Boxplot (interquartile range - IQR) Yöntemi

En sık kullanılan yöntemlerden birisidir. Değişkenin değerleri küçükten büyüğe sıralanır. Çeyreklerine (yüzdeliklerine) yani Q_1, Q_3 değerlerine karşılık gelen değerler üzerinden bir eşik değer hesaplanır ve bu eşik değere göre aykırı değer tanımı yapılır.



Bu grafik minimum değer, maksimum değer, 3. çeyrekteki değer, 1. çeyrekteki değer ve medyana göre oluşturuluyordu.

Burada eşik değerleri hesaplamak için;

$$IQR = 1.5 \times (Q_3 - Q_1)$$

değeri hesaplanır. Bu değer alt eşik değer ve üst eşik değer hesaplamasında aşağıdaki şekilde kullanılır.

$$\text{Alt Eşik Değer} = Q_1 - IQR$$

$$\text{Üst Eşik Değer} = Q_3 + IQR$$

Genellikle 1. veya 4. yaklaşım kullanılır.