EKSİK DEĞERLERİN ELE ALINMA YÖNTEMLERİ

Bir veri analizi gerçekleştirirken ele alınması gereken en önemli konulardan bir tanesi o veri setinde kayıp değerin olup olmadığıdır. İlk başta bu tür verileri ele alırken hangi stratejileri ne tür verilerde uygulayacağımıza karar verebilmek çok büyük bir anlam taşımaktadır. Bunun için temel 2 yaklaşım mevcuttur. İlk olarak doldurma(imputation) ve silme(deletion) çözümlemeleri yapılabilir. Silme yöntemleri de yine veri setine bağlı olarak kolon bazlı, satır bazlı veya eşlerin silinmesi(pairwise) olarak ele alınan popüler yöntemlerden birkaçıdır. Doldurma yöntemleri hem genel modeller hem de zaman serileri için uygulanabilir niteliktedir. Zaman serilerinde sıklıkla kullanılan yöntemlerden bazıları şunlardır: Ortalama, medyan, örneklem ile rastgele doldurma(trend ve mevsimsellik etkileri dışında). Yine zaman serilerinde ele alınan ve sıklıkla kullanılan yöntemlerden biri de verinin trende göre doğrusal regresyonu ve mevsimsellik etkisine göre lineer regresyonu olarak düşünülebilir. Genel modeller için kategorik verilerde çoklu doldurma ya da lojistik regresyon; sürekli değerlerde ortalama, medyan, mod veya lineer regresyon veri yapısına göre ele alınması gereken temel yöntemlerden sadece birkaçıdır.

Burada eksik değer türlerinin ayrımı da çok önem arz etmektedir. Yani eksik değer türünün tamamen rastgele eksik olması, rastgele eksik olmaması veya rastgele eksik olması durumları ayrı ayrı ele alınması gereken konulardır. Bu yöntemleri kısaca açıklayacak olursam tamamen rastgele eksik verilerde bir veri noktasının eksik olması ile veri kümesindeki herhangi bir değerin eksik olması veya gözlemlenmesi arasında bir ilişki yoktur. Eksik veriler, verilerin yalnızca rastgele bir alt kümesidir diyebiliriz. Rastgele eksik verilerde ise bir veri noktasının eksik olma eğiliminin eksik verilerle ilgili olmadığı, ancak gözlemlenen verilerden bazıları ile ilgili olduğu anlamına gelir. Rastgele eksik olmayan verilerde eksik olma olasılığının bizim bilmediğimiz nedenlerle değiştiği anlamına gelir. Esik olan bir değerin eğilimi ile değerleri arasında bir ilişki olduğu anlamına gelir. Yani oluşan bu kayıp verisi bir sebebi veya mekanizması olduğu için oluşmuştur.

Genel olarak bir veri setinde eksik değerler olması durumunda, birden fazla özniteliğimiz varsa değişkenlerden yalnızca biri eksik değerlerden oluşabilir. O zaman bu değişkenin veri modeliyle son derece alakalı olup olmadığına bakılmalıdır. Değilse, değişkeni tamamen kaldırabiliriz. Diğer bir kriter, eksik değerlerin yüzde kaçının kabul edilebilir olduğudur. Eksik değeri, eksik olmayanın ortalaması için (veya kategorikse modu) değiştirmek, belirli bir noktaya kadar makul olabileceğini düşünmekteyim.

Bir eğitim setinde X'in açıklayıcı değişkenler ve Y'nin de hedef değişken olduğunu varsayarak bir modeli eğittiğimizde ve bunu bir test setine uyguladığınızda, her iki setin de aynı dağılımdan geldiği düşünülür. Önceden işlenmiş bir eğitim setinde bir model eğitirsek (eksik değerleri hesaba katarak) bunu aynı şekilde önceden işlenmiş bir test setine gerçek hayatı en iyi şekilde yansıtabilmesi için yerleştirmemiz gerektiğini düşünüyorum.