TechCareer Data Science Bootcamp Final Projesi Raporu

Final projesi konusu olarak kalp rahatsızlığına dair bir araştırma gerçekleştirmek istedim. Bunun için Kaggle üzerinde bulduğum Cardiovascular Diseases Health Dataset veri setini kullandım. Bu veri seti toplamda 22 değişken içeriyor. Bunlardan birisi kalp rahatsızlığı veya kalp krizi geçirip geçirmemek ile ilgili. Diğer 21 değişken içerisinde kolesterol ve tansiyon verileri, sigara ve alkol kullanımı, beslenme ve kiloya dair bilgiler, subjektif sağlık bilgileri ve demografik bilgiler bulunuyor.Bilgiler 2015'te BRFSS kurumu tarafından telefon üzerinden gerçekleştirilmiş.

Öncelikli olarak veri setimi kontrol ettim, halihazırda ön işleme tabi tutulmuş ve temizlenmiş bir veri setiydi. Binary bilgiler de numerik olarak kodlandığı için kodlamaya dair el kitabına dayanarak gerekli açıklamaları notebook'uma ekledim.

Sonrasında dağılım grafiklerini incelemeden önce değişkenlerim hakkın nicel bilgiye sahip olmak adına frekans tablolarını oluşturdum. Subjektif (Self-Reported, kişinin beynaına göre) sağlık bilgisine dair olan PhysHlth ve MentHlth değişkenlerinin dağılımının normal olmadığını fark ettim. Çoğunluk olarak 0 gün sıkıntı hissedildiği belirtilmişti fakat 30 gün sıkıntı hissettiğini belirten kişi sayısının frekansı da 2. ve 3. sırada geliyordu. Bununla beraber ilk 5 cevap içindeki diğer günler 5 gün ve altında sağlık problemi yaşadığını belirtmişti. Yani bu durum da kısmen binary bir hali gösteriyor. İnsanlar ya problem pek hissetmiyor ya da hep problem hissediyor gibi bir sonuca varabiliriz. Buna ek olarak da 15 gün bildiren kişi sayısı da ilk 10 değer arasında girmişti. Yani subjektif sağlık değerlendirmesinin bir genelleme üzerine rapor edildiğini söyleyebiliriz.

Demografik durumlara baktığımızda ise yaşlıların çoğunlukta olduğunu ve en çok yığılmanın 60 yaş civarında yaşandığını görüyoruz. Bu noktada yaş dağılımı normale yakın bir grafik oluşturmuştu. Eğitim ve gelir düzeyine bakınca ise beklenmedik şekilde en yüksek sayının üniversite üzeri eğitimli ve 75 bin Dolar üzeri gelir seviyesine sahip olduğu görülüyor. Bu noktalara dayanarak çıkarabiliriz ki örneklem doğrudan popülasyonu temsil edecek nitelikte değil. Bunun nedenleri arasında ise denek seçimindeki zorunlu olarak oluşan sapmalardan bahsedebiliriz. Nedeni ise muhtemelen anketlerin telefonla yapılıyor olmasından dolayıdır. Anket oldukça uzun, ve gün içinde ankete cevap vermeyi kabul eden kişilerin de buna ayıracak zamanı olmalı. Dolayısıyla çalışan ve genç kesim ankete çok az katılım göstermiş. Ekonomik düzeyin de telefonla ulaşılabilirlik konusunda bir koşul barındırdığı yorumuna varabiliriz.

Bu analizimin devamında ise bivariate analizi ile kalp rahatsızlığı durumuna göre gruplandırma yaparak grafikleri ortaya koydum. Kolesterol ve tansiyon ile daha yüksek gözükmekle beraber, eğitim ve gelir durumuna göre de daha yüksek sosyo-ekonomik durumda olanların kalp krizi geçirme oranlarının daha az olduğunu gördüm. Buna paralel olarak subjektif sağlık verileri ve sosyo-ekonomik verileri regresyon analizine koyduğumda da yüksek sos-eko seviyenin daha düşük sağlık problemi rapor ettiği görülüyor. Bu grafik incelerimi takiben de korelasyon tablosunu çıkardım, ve korelasyon durumları çok yüksek değildi. Kalp rahatsızlığı durumu diğer kategorilerle en fazla 0.26 ile subjektif GenHlth (Genel Sağlık Problemi, 5 en yüksek) raporuyla ilişki halindeydi.

Son olarak da modelimi binary classification yapabilecek bir makine öğrenmesi algoritmasıyla eğitmek için gerekli modellere baktım. En uygun modelin Random Forest olacağına karar verdim. Alternatif olarak denediğim logistic regression ile de benzer sonuçlara ulaştım. Tahmin başarısı %55 üzerine çıkmıyor. Algoritma, train setine overfitting yapmaya meyilli ve n_estimators miktarının 30-200 arasında değişmesi fark yaratmıyor.

Daha sonrasında geliştirmek adına feature'ları karşılaştırmak için feature importance fonksiyonunu çağırdım. En yüksek ilişki BMI değeri iler gözükürken, sonrakilerde ise sağlık problemi olan gün bildirimi bilgileri geliyordu. Feature ayarlaması için daha fazla basamak gitmeden çalışmamı burada sonlandırdım.

Orhan Akpınar