# TechCareer DataScience Bootcamp Fest'4 Mezuniyet Projesi Raporu

**Gamze Gezgin**

Çalışmamda, iki sınıflı bir sınıflandırma problemi üzerinde çalıştık. Veriler üzerinde eksik değer kontrolü ve ölçeklendirme işlemleri yapıldı. K-Nearest Neighbors (KNN) ve Logistic Regression modelleri kullanıldı. Çapraz doğrulama ile en iyi hiperparametreler belirlendi. Logistic regressionda veri dengesizliğini çözmek için class weight manuel olarak kullanıldı.

KNN modeli (k=29) ve Logistic Regression ile elde edilen sonuçlar karşılaştırıldığında, her iki modelin de doğruluk oranı (%71-72) benzer seviyede, ancak performans açısından farklılıklar bulunmaktadır. KNN modeli, 1 sınıfında daha yüksek duyarlılık (%94.3) göstererek pozitif sınıfı doğru tahmin etmede daha iyi sonuç verirken, yanlış pozitif tahminler nedeniyle kesinlik (%73.07) daha düşük kalmıştır. Logistic Regression ise daha dengeli bir performans sergileyerek, 0 sınıfını KNN'ye göre daha iyi ayırt etmiş, ancak 1 sınıfında KNN kadar yüksek bir duyarlılık sağlayamamıştır. Karmaşıklık matrislerinden görüldüğü üzere, KNN modelinde 0 sınıfı ciddi şekilde ihmal edilirken (10 doğru tahmin, 49 yanlış pozitif), Logistic Regression bu dengeyi biraz daha iyi sağlamıştır. Sonuç olarak, pozitif sınıfa odaklanılması gereken durumlarda KNN tercih edilebilirken, sınıf dengesi önem taşıyorsa Logistic Regression daha uygun bir seçimdir.

### 1.Veri setinde eksik değerler var mı? Hangi kolonlarda var ve bunlarla nasıl başa çıkılacak?

Null değerler var. Null degerleri bir önceki degerler ile doldurmaya karar verdim, null oranı düşük, satırların diğer özelliklerini kaybetmek istemedigim yani 478 row'u da kaldırmamak için bu yöntemi seçtim.

### 2. "Age", "Credit amount" ve "Duration" değişkenlerinin dağılımları nedir? Aykırı değerler var mı?

Histograma bakarak yapabildiğim yorumlar;

“age”: 24-28 yaş arası kredi başvurusu daha yogun ve bu 30 yaştan sonra genelde düşme eğilimindedir. Bunun sebebi kariyerine yeni başlayan insanların henüz sermaye biriktiremeiş olması olabilir.

“credit amount”: en cok kredi istenen tutar yaklaşık 1000euro, sonrasında giderek azalmış.

“duration”: birim ay olarak ve genelde 12 ay, 24 ay ve 36 ayda yükselmeler var. Bu da 1,2 ve 3 yıllık krediler çekilmesi ile ilişkilidir.

Boxplotlar histogram grafiklerini doğrular niteliktedir. “credit amount” oldukca outlier içeriyor, bunun sebebi cogunlugun büyük oranda 2000 euro civarı olması ve üst sınırlarda para çeken insanların da fazlasıyla bulunması, eger outlierları cıkarırsam bu fazlaca yüksek kredi çeken insanları tahminleme yetenegimi kaybederim. Bu sebepten hiçbir outlier degerine dokunmamaya karar verdim.

### 3. Hedef kolonunda ("iyi" ve "kötü" kredi riski) oran nasıldır? Veri dengesiz mi?

Bu sonuçlara dengeli bir dağılımdır diyemeyiz. %30 risksiz iken %70 riskli grubu oluşturuyor.

### 4. "İyi" kredi riski kategorisindeki bireylerin ortalama "Credit amount" değeri nedir?

Risk değeri kötü olan kişilerin talep ettiği kredi miktarının, risk değeri iyi olan kişilere göre yaklaşık 1000 euro daha az olduğunu görüyoruz.

### 5. "Free" (bedava) konut kategorisindeki bireylerin "Saving accounts" değişkeninin dağılımı nasıldır?

“free” konut kategorisindeki bireylerin büyük çogunlugunun finansal birikimi "little" kategorisindedir.

### 6. "İyi" ve "kötü" kredi riski grupları arasında "Duration" farklılık gösteriyor mu?

Risk teşkil eden grubun duration degeri etmeyenlere göre yaklaşık 6 ay daha kısa.

### 7. Yüksek kredi miktarına sahip bireylerin (75. percentile üzerinde) en sık kullandığı 3 "Purpose" kategorisi nedir?

Yüksek kredi miktarına sahip bireylerin açık bir farkla en çok car, ardından radio/TV ve business amaçlarının olduğunu görüyoruz.