Veri Analizi ve Model Karşılaştırması Raporu

Ali Eren Sezer

Metodoloji

Bu raporda, bir kredi riski tahmin veri seti üzerinde yapılan keşifsel veri analizi (EDA) ve iki farklı makine öğrenmesi modelinin (Lojistik Regresyon ve KNN Regresyon) performans karşılaştırması yer almaktadır. Çalışma şu adımları izlemiştir:

1. **Veri Hazırlığı ve Temizliği**:
    - Veri setindeki eksik değerler kontrol edilip, uygun yöntemlerle (örneğin, ortalama ile doldurma) bu eksiklikler giderilmiştir.
    - Aykırı değerler, IQR (Interquartile Range) yöntemi ile tespit edilip analiz edilmiştir.
2. **Keşifsel Veri Analizi (EDA)**:
    - Yaş, kredi miktarı ve kredi süresi gibi değişkenlerin dağılımları incelenmiş, aykırı değerler tespit edilmiştir.
    - Hedef değişkenin ("Risk") sınıflarındaki dağılım analiz edilerek veri setinin dengesiz olup olmadığı belirlenmiştir.
    - Kredi miktarı 75. percentil üzerinde olan bireylerin en sık kullandığı "Purpose" kategorileri incelenmiştir.
3. **Modelleme**:
    - **Lojistik Regresyon**: Sınıflandırma problemini çözmek için kullanılmıştır.
    - **KNN Regresyon**: Kayıp (regresyon) problemine uygun bir model olarak uygulanmıştır.
    - Her iki modelin doğruluğu, hata oranları ve diğer metrikler kullanılarak performansları değerlendirilmiştir.
    -
4. **Değerlendirme ve Karşılaştırma**:
    - Her iki model için doğruluk, hata oranları (MSE) ve karışıklık matrisi gibi metriklerle performans karşılaştırması yapılmıştır.
    - İstatistiksel analizler (t-testi) ile model performansları arasındaki farklar test edilmiştir.

Cevaplanması Gereken Sorular:

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

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

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

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

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

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

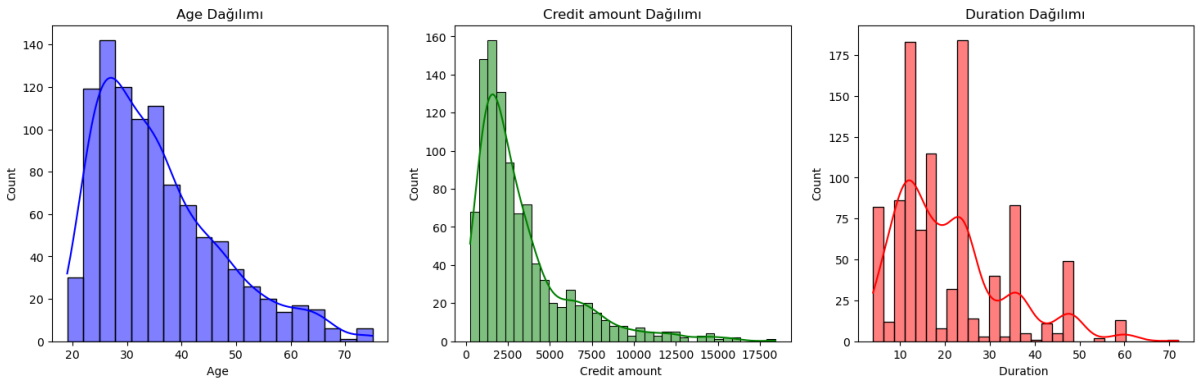
7. Yüksek kredi miktarına sahip bireylerin (75. percentile üzerinde) en sık kullandığı 3 "Purpose" kategorisi nedir?

Cevaplar:

1. Veri setinde eksik değer bulunmamaktadır.

2. Dağılımlar aşağıdaki grafiklerde gösterilmiştir. Verilerde aykırı değerler vardır.

(ChatGpt yardımı kullanılmıştır.)

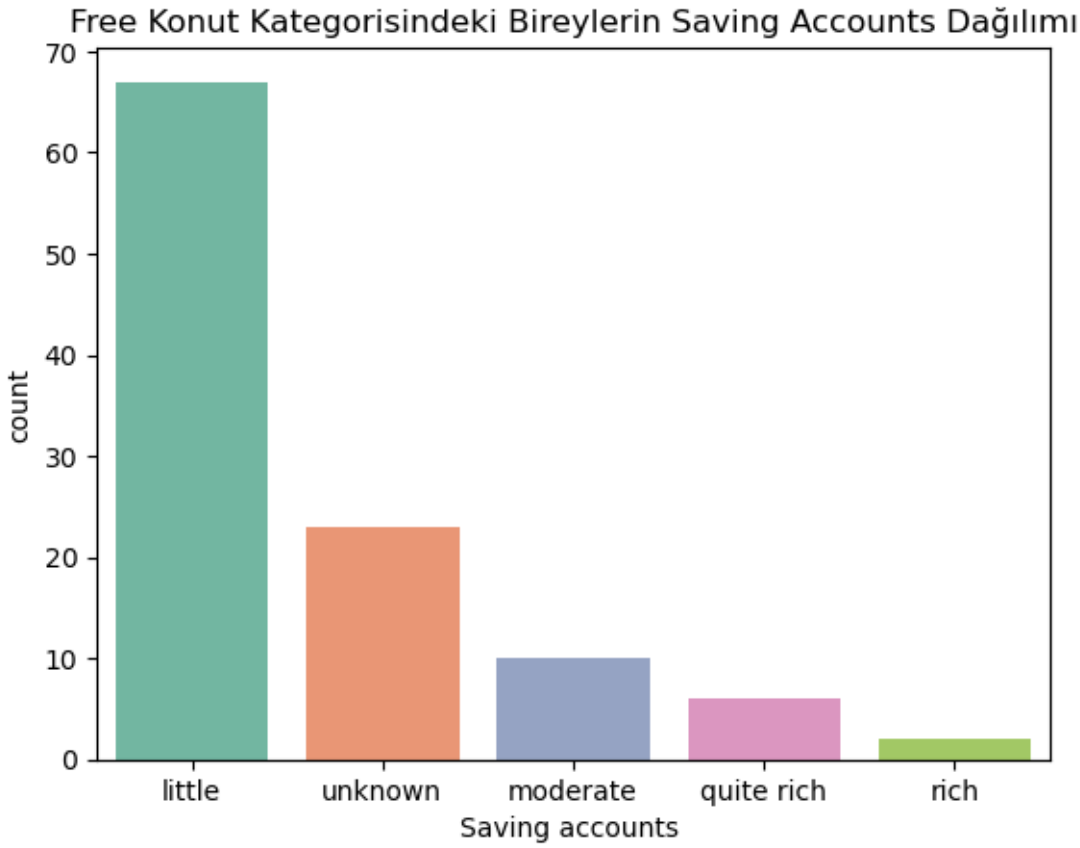


3. İyi Kredi Riski Oranı = %70

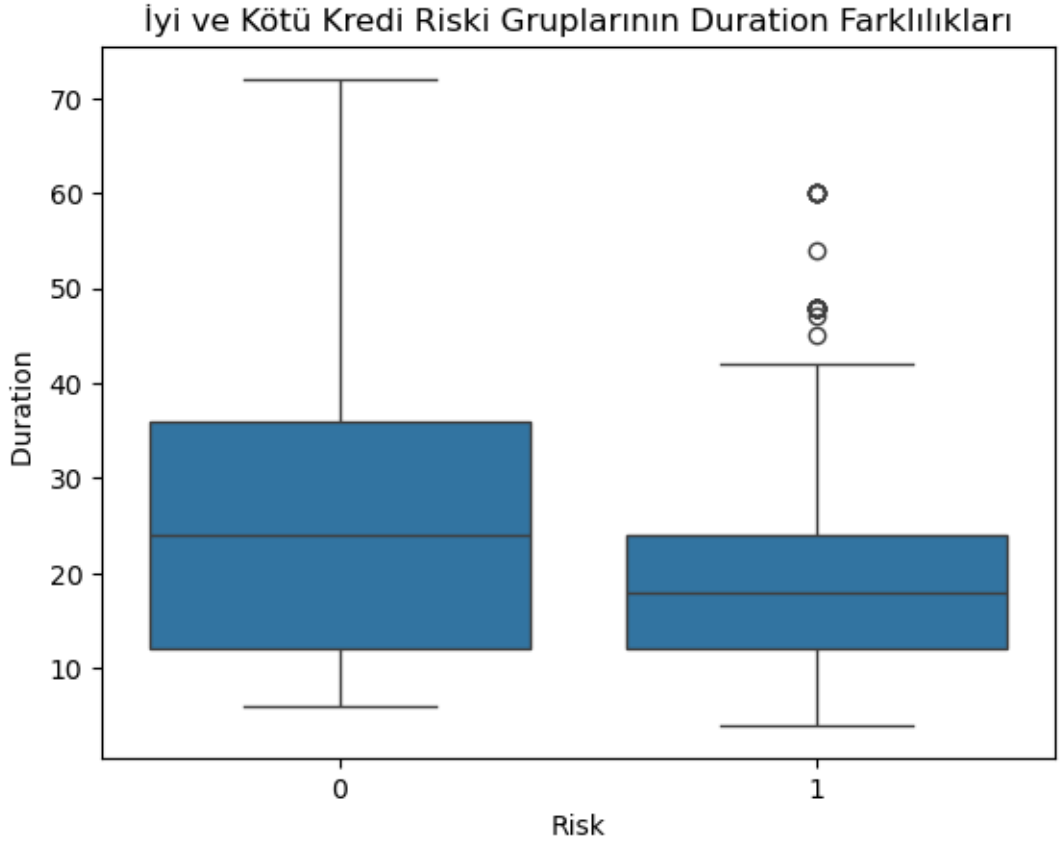
Kötü Kredi Riski Oranı = %30 ve verilerde dengesizlik vardır.

4. İyi Kredi Riski Olan Bireylerin Ortalama Kredi Miktarı: 2985.457142857143

5. "Free" (bedava) konut kategorisindeki bireylerin "Saving accounts" değişkeninin dağılımı aşağıda verilmiştir.



6. Evet,farklılık vardır.Aşağıda verilen grafikte bu farklılık görülmektedir.



7. Yüksek Kredi Miktarına Sahip Bireylerin En Sık Kullandığı 3 Purpose Kategorisi:

Purpose	
car	108
radio/TV	39
business	39

Sonuçlar ve Kilit Bulgular

1. **Eksik Değerler ve Aykırı Değerler**:
    - Veri setindeki eksik değerler, uygun imputation teknikleriyle doldurulmuştur. Aykırı değerler ise IQR yöntemi ile tespit edilmiştir, ancak verinin genel yapısına zarar vermemek adına bu aykırı değerler veri setinden çıkarılmamıştır.
    - Yaş, kredi miktarı ve kredi süresi gibi sayısal değişkenlerde bazı aykırı değerler bulunmuştur. Bu, model performansını etkileyebilecek önemli bir faktör olmuştur.
2. **Veri Dengesizliği**:
    - Hedef değişken olan "Risk" kolonunda "iyi" ve "kötü" kredi riski sınıfları arasında belirgin bir dengesizlik görülmüştür. "Kötü" kredi riski daha az temsil edilmiştir, bu da modelin öğrenmesini zorlaştırabilir.
3. **Model Performansı**:
    - **Lojistik Regresyon**:
        - Modelin doğruluğu %72 olarak hesaplanmıştır.
        - Karışıklık matrisi ve sınıflandırma raporu, modelin "iyi" kredi riski kategorisini doğru tahmin etmede güçlü olduğunu ancak "kötü" kredi riski tahminlerinde bazı zorluklar yaşadığını göstermektedir.
    - **KNN Regresyon**:
        - Modelin hata oranı (MSE) 0.2728 olarak hesaplanmıştır.
        - KNN, regresyon problemi olarak ele alındığından, sürekli değişken tahmini için kullanılmıştır. Kategorik sınıflandırma problemlerinde kullanıldığında sınırlamalar görülebilir.
        -

4. **İstatistiksel Analiz**:
    - İstatistiksel testler (t-testi) uygulandığında, Lojistik Regresyon ve KNN Regresyon modelleri arasında anlamlı bir fark bulunmamıştır. Bu, her iki modelin benzer şekilde performans gösterdiği sonucuna varılmasını sağlamıştır.

Model Karşılaştırması ve Değerlendirme

Bu çalışmada Lojistik Regresyon ve KNN (K-En Yakın Komşu) algoritmalarının performansları karşılaştırılmıştır. Değerlendirme, aşağıdaki temel metriklere dayandırılmıştır:

1. Lojistik Regresyon

- **Precision**: Model, sınıf 1 için %73 gibi yüksek bir doğruluğa sahipken, sınıf 0 için %58 düzeyinde kalmıştır.
- **Recall**: Sınıf 1 için oldukça yüksek bir değer (%94) elde edilmiş, ancak sınıf 0 için bu oran oldukça düşüktür (%19).
- **F1-Score**: Genel F1-Skoru sınıf 1 için %83, sınıf 0 için %28 olarak hesaplanmıştır.
- **Accuracy**: Modelin genel doğruluğu %72'dir.
- Lojistik Regresyon, dengesiz veri setlerinde yüksek sınıflandırma başarısı sunmuş, ancak sınıf 0 için düşük performans sergilemiştir.

2. KNN Classifier

- **Precision**: Sınıf 1 için %71 doğruluk sağlanmış, ancak sınıf 0 için bu oran %31 düzeyindedir.

- **Recall**: Sınıf 1 için %81, sınıf 0 için %20 olarak hesaplanmıştır.
- **F1-Score**: Sınıf 1 için %75, sınıf 0 için %24 olarak belirlenmiştir.
- **Accuracy**: Modelin genel doğruluğu %63 olarak gerçekleşmiştir.
- KNN, sınıf dengesizliği nedeniyle sınıf 0'da daha düşük performans göstermiştir. Bununla birlikte, model sınıf 1 için kabul edilebilir bir performans sunmuştur.

Sonuçların Karşılaştırılması ve İleriye Yönelik Yorumlar

Performans Kriterleri: Lojistik Regresyon, genel doğruluk ve F1-Skor açısından KNN modeline kıyasla daha iyi sonuçlar elde etmiştir. Ancak, sınıf 0 için her iki modelde de yetersiz bir performans gözlemlenmiştir. Bu durum, veri setindeki sınıf dengesizliğinden kaynaklanmaktadır.

Model Seçimi: Sınıf 1'in (örneğin, iyi kredi riski) tahmin edilmesi kritik bir öncelik taşıyorsa, Lojistik Regresyon tercih edilmelidir. Ancak daha dengeli bir model arayışı için, sınıf dengesini sağlamak adına veri artırma (data augmentation) veya sınıf ağırlıklandırma gibi yöntemler uygulanmalıdır.