İçindekiler

[1. Veri Setindeki Değişkenler Hakkında 3](#_Toc199070054)

[2. Veri Setinin Yapısı ve Eksik Değer Analizi 3](#_Toc199070055)

[2.1 Değişkenlerin Genel Yapısı 4](#_Toc199070056)

[2.2 Eksik Değer Durumu 4](#_Toc199070057)

[2.3 Eksik Değerlere Yönelik Yorum 4](#_Toc199070058)

[3. Sayısal Değişkenlere Ait İstatistiksel Özet 5](#_Toc199070059)

[3.1 Açıklayıcı Yorumlar 5](#_Toc199070060)

[4. Hedef Değişken (BAD) Dağılımı 6](#_Toc199070061)

[4.1 Yorum: 6](#_Toc199070062)

[5. Sayısal Değişkenler Arasındaki Korelasyon Analizi 7](#_Toc199070063)

[5.1 Yorumlar 7](#_Toc199070064)

[6. Sayısal Değişkenlerin Dağılım Analizi 8](#_Toc199070065)

[6.1 Yorumlar 8](#_Toc199070066)

[7. Kategorik Değişkenlerin Hedef Değişken (BAD) ile İlişkisi 9](#_Toc199070067)

[7.1 REASON Değişkeni ve BAD Dağılımı 9](#_Toc199070068)

[7.2 JOB Değişkeni ve BAD Dağılımı 10](#_Toc199070069)

[8. Veri Ön İşleme 10](#_Toc199070070)

[8.1 Eksik Verilerin İşlenmesi 10](#_Toc199070071)

[8.2 Kategorik Değişkenlerin Dönüştürülmesi 11](#_Toc199070072)

[8.3 Sayısal Değişkenlerin Ölçeklenmesi 11](#_Toc199070073)

[8.4 Eğitim ve Test Veri Ayırımı 11](#_Toc199070074)

[9. Model Geliştirme – Lojistik Regresyon 11](#_Toc199070075)

[9.1 Model Yapılandırması 11](#_Toc199070076)

[9.2 Model Başarım Değerlendirmesi 12](#_Toc199070077)

[9.3 Sınıflandırma Raporu 12](#_Toc199070078)

[9.4 Confusion Matrix (Karışıklık Matrisi) 13](#_Toc199070079)

[9.5 ROC Eğrisi ve AUC Skoru 14](#_Toc199070080)

[10. Model Geliştirme – Random Forest 14](#_Toc199070081)

[10.1 Model Yapılandırması 14](#_Toc199070082)

[10.2 Model Başarım Değerlendirmesi 14](#_Toc199070083)

[10.3 Sınıflandırma Raporu 15](#_Toc199070084)

[10.4 Confusion Matrix 15](#_Toc199070085)

[10.5 ROC Eğrisi ve AUC Skoru 16](#_Toc199070086)

[11. Model Geliştirme – XGBoost 16](#_Toc199070087)

[11.1 Model Yapılandırması 16](#_Toc199070088)

[11.2 Model Başarım Değerlendirmesi 17](#_Toc199070089)

[11.3 Sınıflandırma Raporu 17](#_Toc199070090)

[11.4 Confusion Matrix 17](#_Toc199070091)

[11.5 ROC Eğrisi ve AUC Skoru 18](#_Toc199070092)

[11.6 Genel Değerlendirme 18](#_Toc199070093)

[12. MODEL KARŞILAŞTIRMA – LOJİSTİK REGRESYON vs RANDOM FOREST vs XGBOOST 19](#_Toc199070094)

[12.1 Model Performans Özeti Tablosu 19](#_Toc199070095)

[12.2 Değerlendirme ve Yorum 19](#_Toc199070096)

[13. Model İyileştirme – XGBoost (RandomizedSearchCV ile Hiperparametre Optimizasyonu) 19](#_Toc199070097)

[13.1 Parametre Aralıkları 20](#_Toc199070098)

[13.2 En İyi Parametreler 20](#_Toc199070099)

[13.3 Model Performans Metrikleri 21](#_Toc199070100)

[13.4 Sınıflandırma Raporu 21](#_Toc199070101)

[13.5 Confusion Matrix 22](#_Toc199070102)

[13.6 ROC Eğrisi 22](#_Toc199070103)

[13.7 Sonuç 22](#_Toc199070104)

[14. Tüm Modellerin Karşılaştırması 23](#_Toc199070105)

[14.1 Performans Karşılaştırma Tablosu 23](#_Toc199070106)

[14.2 Değerlendirme ve Yorum 23](#_Toc199070107)

[15. Sonuçlar 24](#_Toc199070108)

[16. Gözetimsiz Öğrenme – KMeans Kümeleme 24](#_Toc199070109)

[16.1 Uygulanan Adımlar 24](#_Toc199070110)

[16.2 Elbow Yöntemi ile Optimal Küme Sayısı 25](#_Toc199070111)

[16.3 PCA ile 2 Boyutlu Görselleştirme 26](#_Toc199070112)

[16.4 Küme Özelliklerinin Karşılaştırması 26](#_Toc199070113)

# 

# 1. Veri Setindeki Değişkenler Hakkında

| **Değişken Adı** | **Açıklama** |
| --- | --- |
| **BAD** | Hedef değişken. Müşteri borcunu ödemediyse 1, ödediyse 0. (**Kredi riski**) |
| **LOAN** | Talep edilen kredi miktarı (USD cinsinden) |
| **MORTDUE** | Mevcut ipotek borcu (USD cinsinden) |
| **VALUE** | Evin tahmini piyasa değeri (USD cinsinden) |
| **REASON** | Kredinin nedeni (Ev geliştirme – HomeImp / Borç transferi – DebtCon) |
| **JOB** | Müşterinin işi/mesleği (örn. Öğretmen, Ofis çalışanı, Mavi yaka vs.) |
| **YOJ** | Aynı işte geçirilen yıl sayısı |
| **DEROG** | Geçmişteki olumsuz kredi kayıtlarının sayısı (negatif durumlar) |
| **DELINQ** | Son zamanlardaki ödeme gecikmelerinin sayısı |
| **CLAGE** | En eski kredi hesabının kaç aydır açık olduğu (kredi yaşı) |
| **NINQ** | Son 6 ayda yapılan kredi başvuru sayısı |
| **CLNO** | Toplam açık kredi hesabı sayısı |
| **DEBTINC** | Borç/gelir oranı (Debt-to-Income Ratio) |

# 2. Veri Setinin Yapısı ve Eksik Değer Analizi

Bu çalışmada kullanılan HMEQ veri seti, toplam 5960 gözlem (satır) ve 13 değişkenden (sütun) oluşmaktadır. Değişkenlerden 9’u kesikli/sürekli sayısal (float64), 2’si tam sayı (int64) ve 2’si kategorik (object) türündedir. Hedef değişken, BAD adlı 0/1 formatındaki ikili sınıflandırma etiketidir.

## 2.1 Değişkenlerin Genel Yapısı

Veri seti, müşterilere ait kredi, gelir, borç ve istihdam gibi özelliklerin yanı sıra, kredi geçmişine dair değişkenleri de içermektedir. Bu yönüyle kredi riski modellemeleri için uygundur.

## 2.2 Eksik Değer Durumu

Veri setinde birçok değişkende eksik değer bulunduğu tespit edilmiştir. Aşağıda, eksik değer içeren değişkenler ve eksiklik oranları tablo halinde sunulmuştur:

| **Değişken** | **Eksik Değer Sayısı** | **Eksik Yüzdesi (%)** |
| --- | --- | --- |
| DEBTINC | 1267 | 21.26 |
| DEROG | 708 | 11.88 |
| DELINQ | 580 | 9.73 |
| MORTDUE | 518 | 8.69 |
| YOJ | 515 | 8.64 |
| NINQ | 510 | 8.56 |
| CLAGE | 308 | 5.17 |
| JOB | 279 | 4.68 |
| REASON | 252 | 4.22 |
| CLNO | 222 | 3.72 |
| VALUE | 112 | 1.88 |

## 2.3 Eksik Değerlere Yönelik Yorum

* Özellikle DEBTINC değişkenindeki eksiklik oranı (%21.26), modelleme sürecinde dikkat edilmesi gereken önemli bir duruma işaret etmektedir. Bu kadar yüksek oranlı eksiklik, doğrudan model performansını etkileyebilir.
* DEROG, DELINQ, MORTDUE, YOJ, NINQ, CLAGE gibi değişkenlerdeki eksik oran %5’in üzerinde olup, standart dolgu tekniklerinden daha dikkatli stratejiler gerektirmektedir.
* JOB ve REASON gibi kategorik değişkenlerde eksik değer oranları düşük olmakla birlikte (%4 civarı), uygun bir şekilde kategorik doldurma (örneğin “Unknown” etiketi veya mod değeri ile) yapılmalıdır.

# 3. Sayısal Değişkenlere Ait İstatistiksel Özet

Aşağıda, HMEQ veri setindeki sayısal değişkenlerin merkezi eğilim ve dağılım ölçüleri yer almaktadır. Bu özet, değişkenlerin yapısını anlamaya ve potansiyel aykırı değerleri tespit etmeye yardımcı olmaktadır.

| **Değişken** | **Gözlem Sayısı** | **Ortalama** | **Standart Sapma** | **Min** | **25%** | **Medyan (50%)** | **75%** | **Maksimum** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| BAD | 5960 | 0.199 | 0.399 | 0 | 0 | 0 | 0 | 1 |
| LOAN | 5960 | 18608 | 11207 | 1100 | 11100 | 16300 | 23300 | 89900 |
| MORTDUE | 5442 | 73761 | 44458 | 2063 | 46276 | 65019 | 91488 | 399550 |
| VALUE | 5848 | 101776 | 57386 | 8000 | 66075 | 89236 | 119824 | 855909 |
| YOJ | 5445 | 8.92 | 7.57 | 0 | 3 | 7 | 13 | 41 |
| DEROG | 5252 | 0.25 | 0.85 | 0 | 0 | 0 | 0 | 10 |
| DELINQ | 5380 | 0.45 | 1.13 | 0 | 0 | 0 | 0 | 15 |
| CLAGE | 5652 | 179.77 | 85.81 | 0 | 115.12 | 173.47 | 231.56 | 1168.23 |
| NINQ | 5450 | 1.19 | 1.73 | 0 | 0 | 1 | 2 | 17 |
| CLNO | 5738 | 21.30 | 10.14 | 0 | 15 | 20 | 26 | 71 |
| DEBTINC | 4693 | 33.78 | 8.60 | 0.52 | 29.14 | 34.82 | 39.00 | 203.31 |

## 3.1 Açıklayıcı Yorumlar

* **BAD** değişkeninin ortalaması 0.199, yani gözlemlerin yaklaşık %20’si borcunu ödeyememiştir. Bu da veri setinde sınıf dengesizliği olabileceğini gösterir.
* **LOAN**, **MORTDUE** ve **VALUE** gibi finansal değişkenlerde yüksek standart sapma gözlenmiştir. Bu durum, müşteri profilleri arasında ciddi farklılıklar olduğunu ve bazı **aykırı değerlerin** olabileceğini gösterir.
* **YOJ (İş deneyimi yılı)** değişkeninin ortalaması 9 yıl civarındadır; ancak 0 olan gözlemler de mevcuttur, bu da bazı müşterilerin işsiz olabileceğine işaret eder.
* **DEROG** ve **DELINQ**, genellikle sıfır çevresinde yoğunlaşan ve sağa çarpık dağılıma sahip değişkenlerdir. Maksimum değerlerin (10 ve 15) oldukça yüksek olması dikkat çekicidir.
* **CLAGE (kredi geçmişi yaşı)** değişkeni 179 ay (yaklaşık 15 yıl) ortalamaya sahiptir. Ancak maksimum değerin 1168 ay (yaklaşık 97 yıl) olması, ekstrem aykırı değerlere işaret eder.
* **DEBTINC (borç/gelir oranı)** değişkeninin ortalaması %33,78’dir. %200 üzeri maksimum değer, yüksek borçluluk durumuna sahip bazı ekstrem örneklerin bulunduğunu göstermektedir.

# 4. Hedef Değişken (BAD) Dağılımı

Aşağıdaki grafik, BAD değişkeninin sınıf dağılımını göstermektedir. Bu değişken, müşterinin kredi borcunu ödeyip ödemediğini temsil eder. BAD = 0 müşterinin borcunu ödediğini, BAD = 1 ise borcunu ödeyemediğini ifade eder.

metin, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

## 4.1 Yorum:

Grafikte görüldüğü üzere:

* BAD = 0 sınıfına ait müşteri sayısı yaklaşık 4800 civarındadır.
* BAD = 1 sınıfına ait müşteri sayısı ise yaklaşık 1150 civarındadır.
* Bu durum, sınıflar arasında belirgin bir dengesizlik olduğunu göstermektedir. Yaklaşık olarak müşterilerin %80’i borcunu ödemiş, %20’si ise ödeyememiştir.

# 5. Sayısal Değişkenler Arasındaki Korelasyon Analizi

Aşağıdaki korelasyon matrisi, veri setindeki sayısal değişkenlerin birbiriyle olan doğrusal ilişkilerini göstermektedir. Değişkenler arasındaki korelasyon katsayıları -1 ile +1 arasında değer alır; +1 mükemmel pozitif ilişkiyi, -1 ise mükemmel negatif ilişkiyi ifade eder. 0’a yakın değerler ise zayıf ilişkiyi temsil eder.

ekran görüntüsü, metin, kare, kalıp, desen, düzen içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

## 5.1 Yorumlar

* **MORTDUE** ile **VALUE** arasında oldukça yüksek pozitif korelasyon bulunmaktadır (0.88). Bu, evin değeri arttıkça ipotek borcunun da genellikle arttığını göstermektedir.
* Hedef değişken olan **BAD** ile en güçlü ilişkiler:
  + **DELINQ (0.35)**: Geciken ödeme sayısı arttıkça kredi riski de artmaktadır.
  + **DEROG (0.28)**: Geçmişteki olumsuz kredi kayıtları BAD değişkeniyle orta düzeyde ilişkilidir.
  + **DEBTINC (0.20)** ve **NINQ (0.17)** gibi değişkenlerin de kredi riskiyle zayıf ama dikkate değer ilişkileri vardır.
* **CLAGE** değişkeni (kredi geçmişi süresi), BAD ile negatif ilişkilidir (-0.17). Yani, kredi geçmişi daha uzun olan bireylerin borcunu ödeme olasılığı daha yüksektir.
* **LOAN** ve **BAD** arasında anlamlı bir ilişki gözlemlenmemektedir (korelasyon ≈ -0.08).

# 6. Sayısal Değişkenlerin Dağılım Analizi

Aşağıdaki grafikler, veri setindeki sayısal değişkenlerin dağılımlarını göstermektedir. Histogramlar, değişkenlerin değer aralıklarına göre gözlem sayısını görselleştirerek, dağılım biçimi ve olası aykırı değerler hakkında bilgi verir.

diyagram, metin, plan, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

## 6.1 Yorumlar

* BAD değişkeni ikili (binary) yapıda olduğu için dağılımı iki sütundan oluşmaktadır. Sınıflar arasında ciddi bir dengesizlik olduğu gözlemlenmektedir.
* LOAN, MORTDUE, VALUE gibi parasal değişkenlerde **sağa çarpık dağılım** (right-skewed) hakimdir. Bu durum, veri setinde büyük kredi taleplerinin az sayıda müşteriyle sınırlı olduğunu gösterir. Özellikle VALUE ve MORTDUE değişkenlerinde **yüksek uçlu aykırı değerler** gözlemlenmektedir.
* DEROG, DELINQ, NINQ gibi geçmiş kredi problemi veya başvuru sayısını gösteren değişkenlerde gözlemlerin çoğu **0 ile 2 arasında yoğunlaşmıştır**, yani büyük çoğunlukta sorunlu kredi geçmişi bulunmamaktadır.
* CLAGE (kredi geçmişi süresi), yaklaşık **normal dağılıma yakın** olmakla birlikte sağ kuyruğa sahiptir. Maksimum değerin çok yüksek olması dikkat çekicidir.
* DEBTINC (borç/gelir oranı) değişkeni büyük oranda 20–50 arasında yoğunlaşmaktadır; ancak 200’e kadar uzanan uç değerler mevcuttur. Bu durum aşırı borç yükü taşıyan bireylerin varlığına işaret eder.

# 7. Kategorik Değişkenlerin Hedef Değişken (BAD) ile İlişkisi

## 7.1 REASON Değişkeni ve BAD Dağılımı

REASON değişkeni, müşterinin kredi başvuru sebebini ifade etmektedir. İki kategori mevcuttur:

* HomeImp: Ev iyileştirme (renovasyon, tadilat vb.)
* DebtCon: Borç birleştirme/kapatma

metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Gözlemler:**

* Kredi başvuru nedeni DebtCon olan bireyler arasında BAD = 1 oranı (borcunu ödeyememe) daha yüksektir.
* HomeImp grubunda ise hem toplam gözlem sayısı hem de başarısız kredi oranı daha düşüktür.

## 7.2 JOB Değişkeni ve BAD Dağılımı

JOB değişkeni, müşterilerin meslek gruplarını ifade etmektedir. Kategoriler: Mgr, Office, Sales, Self, Other, ProExe.

metin, ekran görüntüsü, çizgi, yazı tipi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Gözlemler:**

* En büyük grup “Other” olup, bu gruptaki BAD = 1 oranı görece yüksektir.
* “Mgr” (yönetici) ve “ProExe” (profesyonel/uzman) gibi mesleklerde borcunu ödeyemeyen müşteri oranı daha düşüktür.
* “Sales” ve “Self” gibi gelir değişkenliği yüksek meslek gruplarında da BAD oranı görece yüksektir.

# 8. Veri Ön İşleme

Modelleme sürecinden önce, veri setinin kullanılabilir hale getirilmesi amacıyla çeşitli ön işleme adımları gerçekleştirilmiştir. Bu adımlar eksik verilerin yönetimi, kategorik değişkenlerin dönüştürülmesi, sayısal değişkenlerin ölçeklenmesi ve eğitim/test veri ayrımını kapsamaktadır.

## 8.1 Eksik Verilerin İşlenmesi

**Sayısal Değişkenler:**

Eksik değerlere sahip sayısal sütunlar (örneğin: DEBTINC, MORTDUE, VALUE, CLAGE) klasik yöntemlerle doldurulmak yerine, **model tabanlı tahmin yaklaşımı** olan **Iterative Imputer** kullanılarak doldurulmuştur.

Bu yöntem, eksik değer içeren her sütunu diğer sütunlara bağlı olarak tahmin eder. Varsayılan olarak BayesianRidge regresyon modeli kullanılmıştır. Böylece verinin dağılımı korunmuş ve yapay bozulmalar engellenmiştir.

**Kategorik Değişkenler:**

REASON ve JOB değişkenlerinde yer alan eksik veriler, "Unknown" etiketi ile doldurulmuştur. Bu yaklaşım, eksikliği bir bilgi olarak modele yansıtmakta ve veri kaybını önlemektedir.

## 8.2 Kategorik Değişkenlerin Dönüştürülmesi

Kategorik değişkenler, makine öğrenmesi algoritmaları tarafından işlenebilmesi için **etiket kodlaması (Label Encoding)** yöntemi ile sayısal forma dönüştürülmüştür. Bu işlem sırasında:

* REASON (kredi başvuru nedeni)
* JOB (meslek grubu)

değişkenleri ayrı ayrı encode edilmiştir.

## 8.3 Sayısal Değişkenlerin Ölçeklenmesi

Sayısal veriler arasındaki farklı ölçek problemini ortadan kaldırmak için, tüm sayısal sütunlar StandardScaler kullanılarak **standart normal dağılıma** dönüştürülmüştür (ortalama = 0, standart sapma = 1). Bu işlem, özellikle mesafe tabanlı algoritmalar (k-NN, SVM) ve regresyon modelleri için performans iyileştirmesi sağlamaktadır.

## 8.4 Eğitim ve Test Veri Ayırımı

Son olarak, modelin eğitim ve test aşamalarında farklı veri gruplarıyla değerlendirilmesi amacıyla veri seti:

* %80 eğitim
* %20 test

olacak şekilde, rastgele ve sınıf oranlarını koruyarak (stratify=y) ayrılmıştır.

# 9. Model Geliştirme – Lojistik Regresyon

Bu bölümde, kredi riski tahmini için temel ve yorumlanabilir bir model olan **Lojistik Regresyon** uygulanmıştır. Model, eksik veri işleme ve ön işlem adımları tamamlandıktan sonra eğitilmiş; başarımları ise çeşitli sınıflandırma metrikleri ve grafiklerle değerlendirilmiştir.

## 9.1 Model Yapılandırması

* **Model:** LogisticRegression (sklearn)
* **Parametreler:**
  + max\_iter=1000 – iterasyon artırılarak yakınsama sağlanmıştır.
  + class\_weight='balanced' – sınıf dengesizliğine karşı ağırlıklandırma yapılmıştır.
* **Eğitim/Test Ayrımı:** %80 eğitim – %20 test, stratify=y ile dengeli dağılım korunmuştur.
* **Ön İşleme:**
  + Sayısal eksik veriler IterativeImputer ile model-tabanlı doldurulmuştur.
  + Kategorik eksik veriler "Unknown" etiketiyle doldurulmuştur.
  + LabelEncoder ile kategorik veriler sayısallaştırılmıştır.
  + Sayısal veriler StandardScaler ile normalize edilmiştir.

## 9.2 Model Başarım Değerlendirmesi

Aşağıda, modelin test verisi üzerindeki başarımı temel sınıflandırma metrikleri ile sunulmaktadır:

| **Metrik** | **Değer** | **Açıklama** |
| --- | --- | --- |
| Accuracy | 0.737 | Genel doğruluk oranı |
| Precision | 0.391 | Riskli sınıf için doğru pozitif tahmin oranı |
| Recall | 0.567 | Riskli sınıfın yakalanma oranı |
| F1 Skoru | 0.463 | Precision ve Recall’un dengeli ortalaması |
| ROC AUC | 0.761 | Sınıflar arası ayrım gücü (ideal: 1.0) |

Bu değerlere göre modelin genel doğruluğu %73,7’dir. Ancak daha kritik olan, **riskli müşterileri (BAD = 1)** tespit etme yeteneğidir. Model bu sınıfı %56,7 oranında yakalamaktadır ki bu, finansal sistemlerde kabul edilebilir düzeyde bir duyarlılıktır. Bununla birlikte, precision düşük olduğu için model bazı müşterileri gereksiz yere riskli olarak tahmin etmiştir (false positive).

## 9.3 Sınıflandırma Raporu

| **Sınıf** | **Precision** | **Recall** | **F1 Skoru** | **Support** |
| --- | --- | --- | --- | --- |
| 0 – Borcunu ödeyen | 0.88 | 0.78 | 0.83 | 954 |
| 1 – Borcunu ödeyemeyen | 0.39 | 0.57 | 0.46 | 238 |

* **Pozitif sınıf (BAD = 1)** için F1 skoru %46,3 olup model bu grubu ayırt etmekte kısmen başarılıdır.
* Negatif sınıf için precision yüksek olmasına rağmen recall düşüktür; yani model “ödeyecek” dediği kişilerin çoğu doğru tahmin edilmiştir.
* Bu sonuçlar, **Recall > Precision** stratejisinin daha uygun olduğunu göstermektedir.

## 9.4 Confusion Matrix (Karışıklık Matrisi)

**metin, ekran görüntüsü, diyagram, yazı tipi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

* **744** doğru negatif, **135** doğru pozitif tahmin yapılmıştır.
* **103** riskli müşteri atlanmış (false negative), **210** risksiz müşteriye yanlış alarm verilmiştir (false positive).

## 9.5 ROC Eğrisi ve AUC Skoru

**metin, diyagram, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

* ROC eğrisi altında kalan alan (AUC): **0.76**  
  Bu değer, modelin pozitif ve negatif sınıfları ayırt etme başarısının iyi düzeyde olduğunu göstermektedir. Eğri, ideal olan sol üst köşeye yakın seyretmektedir.

# 10. Model Geliştirme – Random Forest

Lojistik Regresyon modeli ile elde edilen sınırlı başarıdan sonra, daha güçlü ve esnek bir algoritma olan **Random Forest Classifier** uygulanmıştır. Bu model, karar ağaçlarının topluluğu olarak çalışır ve karmaşık ilişkileri daha iyi öğrenebilme kapasitesine sahiptir.

## 10.1 Model Yapılandırması

* **Model:** RandomForestClassifier (sklearn)
* **Parametreler:** Varsayılan değerler, random\_state=42, class\_weight='balanced'
* **Eğitim/Test Oranı:** %80 / %20
* **Veri Ön İşleme:** Lojistik regresyon ile aynıdır (imputation, encoding, scaling uygulanmıştır)

## 10.2 Model Başarım Değerlendirmesi

Modelin test verisi üzerindeki genel başarı durumu aşağıdaki metriklerle özetlenmiştir:

| **Metrik** | **Değer** | **Açıklama** |
| --- | --- | --- |
| Accuracy | 0.912 | Genel doğruluk oranı |
| Precision | 0.946 | Riskli sınıf tahminlerinin doğruluğu |
| Recall | 0.597 | Gerçek riskli müşterilerin yakalanma oranı |
| F1 Skoru | 0.732 | Precision ve Recall’un dengeli ortalaması |
| ROC AUC | 0.964 | Sınıflar arası ayrım gücü (çok yüksek) |

**Açıklama:** Modelin doğruluk ve AUC skoru oldukça yüksek seviyelerdedir. Özellikle ROC AUC skoru, Random Forest modelinin sınıfları birbirinden güçlü bir biçimde ayırabildiğini göstermektedir.

## 10.3 Sınıflandırma Raporu

| **Sınıf** | **Precision** | **Recall** | **F1 Skoru** | **Support** |
| --- | --- | --- | --- | --- |
| 0 – Borcunu ödeyen | 0.91 | 0.99 | 0.95 | 954 |
| 1 – Borcunu ödeyemeyen | 0.95 | 0.60 | 0.73 | 238 |

* **Precision çok yüksek** → Model "riskli" dediği kişilerin çoğunu doğru tahmin ediyor.
* **Recall (0.60)** → Her ne kadar daha yüksek istense de, lojistik regresyona göre ciddi bir gelişme var.
* F1 skoru (0.73) → Sınıflar arası denge daha başarılı.

## 10.4 Confusion Matrix

**metin, ekran görüntüsü, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

* Doğru tahmin oranı çok yüksektir. Sadece 8 risksiz kişi yanlış şekilde riskli sınıfa atanmıştır.
* Riskli olduğu hâlde kaçırılan (false negative) müşteri sayısı 96’dır — bu, lojistik regresyona göre biraz daha fazla olsa da, model genel anlamda daha yüksek **güvenilirlikte** çalışmaktadır.

## 10.5 ROC Eğrisi ve AUC Skoru

**metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

* ROC eğrisinin altında kalan alan (AUC): **0.964**
* Bu değer, Random Forest modelinin sınıfları neredeyse hatasız bir şekilde ayırt ettiğini göstermektedir.  
  Grafik ideal eğri olan sol üst köşeye oldukça yakındır.

# 11. Model Geliştirme – XGBoost

Bu bölümde kredi riski tahmini için **XGBoost (Extreme Gradient Boosting)** algoritması uygulanmış ve sonuçları diğer modellerle karşılaştırılmıştır. Boosting tabanlı bu model, özellikle karmaşık ilişkileri öğrenme kabiliyetiyle öne çıkar.

## 11.1 Model Yapılandırması

* **Model:** XGBClassifier (xgboost)
* **Parametreler:** eval\_metric='logloss', random\_state=42
* **Veri İşleme:** Lojistik Regresyon ve Random Forest modelleriyle aynı yapı kullanılmıştır.
* **Eğitim/Test oranı:** %80 / %20 stratified split

## 11.2 Model Başarım Değerlendirmesi

| **Metrik** | **Değer** | **Açıklama** |
| --- | --- | --- |
| Accuracy | 0.918 | Genel doğruluk oranı |
| Precision | 0.912 | Riskli tahminlerin doğruluk oranı |
| Recall | 0.651 | Gerçek riskli müşterilerin yakalanma oranı |
| F1 Skoru | 0.760 | Precision ve Recall’un dengeli ortalaması |
| ROC AUC | 0.946 | Sınıflar arası ayrım gücü (yüksek) |

Modelin doğruluğu yüksek, AUC değeri oldukça güçlüdür. Recall oranı önceki modellere göre iyileşmiş ve dengeli bir F1 skoru elde edilmiştir.

## 11.3 Sınıflandırma Raporu

| **Sınıf** | **Precision** | **Recall** | **F1-score** | **Support** |
| --- | --- | --- | --- | --- |
| 0 – Borcunu ödeyen | 0.92 | 0.98 | 0.95 | 954 |
| 1 – Borcunu ödeyemeyen | 0.91 | 0.65 | 0.76 | 238 |

* Model, riskli müşterileri (%65) oranında yakalayabilmiş, bu da kredi riski için **önemli ve kabul edilebilir bir performanstır.**
* F1 skorundaki artış (0.76), dengeli başarıyı temsil etmektedir.

## 11.4 Confusion Matrix

**metin, ekran görüntüsü, diyagram, yazı tipi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

* **155** riskli müşteri başarıyla tespit edilmiştir.
* Modelin false positive sayısı düşüktür (sadece 15 müşteri yanlışlıkla riskli sınıfa atanmıştır).

## 11.5 ROC Eğrisi ve AUC Skoru

**metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

* ROC eğrisi, ideal eğri olan sol üst köşeye oldukça yakındır.
* **AUC skoru: 0.95**, bu da modelin sınıflar arasında güçlü bir ayrım yaptığını gösterir.

## 11.6 Genel Değerlendirme

XGBoost modeli, diğer modellere kıyasla:

* En dengeli Precision-Recall oranını,
* En güçlü F1 skorunu üretmiştir.

Bu nedenle kredi riski tahmin problemi için XGBoost, **en başarılı model olarak öne çıkmaktadır.**  
Hiperparametre ayarlamaları ve özellik önem analizleriyle daha da iyileştirilebilir.

# 12. MODEL KARŞILAŞTIRMA – LOJİSTİK REGRESYON vs RANDOM FOREST vs XGBOOST

## 12.1 Model Performans Özeti Tablosu

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **ROC AUC** |
| --- | --- | --- | --- | --- | --- |
| **Lojistik Regresyon** | 0.737 | 0.391 | 0.567 | 0.463 | 0.761 |
| **Random Forest** | 0.913 | 0.947 | 0.597 | 0.732 | 0.964 |
| **XGBoost** | 0.918 | 0.912 | 0.651 | 0.760 | 0.946 |

## 12.2 Değerlendirme ve Yorum

* **Accuracy** açısından en başarılı model: XGBoost (%91.8)
* **Recall** açısından en başarılı model: XGBoost (%65.1) – riskli müşteriyi yakalamada önemli
* **F1 Score** açısından: XGBoost > RF > LR (denge açısından üstünlük)
* **ROC AUC** açısından: Random Forest > XGBoost > LR

Riskli müşterileri mümkün olduğunca yakalamak için **XGBoost daha dengeli ve güvenilir bir modeldir.**

Random Forest AUC açısından daha iyi ama XGBoost daha az false positive verir.  
Lojistik regresyon ise zayıf yorumlayabilmektedir.

# 13. Model İyileştirme – XGBoost (RandomizedSearchCV ile Hiperparametre Optimizasyonu)

Bu bölümde XGBoost modelinin başarımını artırmak amacıyla geniş parametre aralığına sahip **RandomizedSearchCV** yöntemi kullanılmış ve model 3 katlı çapraz doğrulama (CV=3) ile optimize edilmiştir. Amaç, özellikle F1 skoru ve recall oranını artırarak riskli müşteri tespitinde daha başarılı sonuçlar elde etmektir.

## 13.1 Parametre Aralıkları

| **Parametre** | **Denenen Aralık** |
| --- | --- |
| n\_estimators | [100, 200, 300] |
| max\_depth | [3, 5, 7, 10] |
| learning\_rate | [0.01, 0.05, 0.1] |
| subsample | [0.6, 0.8, 1.0] |
| colsample\_bytree | [0.6, 0.8, 1.0] |
| gamma | [0, 1, 5] |
| min\_child\_weight | [1, 3, 5] |
| reg\_alpha | [0, 0.1, 0.5] |
| reg\_lambda | [1, 1.5, 2] |

Toplam 50 rastgele kombinasyon denenmiş ve f1 skoru hedeflenerek en iyi model belirlenmiştir.

## 13.2 En İyi Parametreler

| **Parametre** | **Seçilen Değer** |
| --- | --- |
| n\_estimators | 300 |
| max\_depth | 10 |
| learning\_rate | 0.1 |
| subsample | 1.0 |
| colsample\_bytree | 0.6 |
| gamma | 0 |
| min\_child\_weight | 1 |
| reg\_alpha | 0.5 |
| reg\_lambda | 1 |

Eğitim süresi: **11.77 saniye’dir**

## 13.3 Model Performans Metrikleri

| **Metrik** | **Değer** |
| --- | --- |
| Accuracy | 0.928 |
| Precision | 0.958 |
| Recall | 0.668 |
| F1 Score | 0.787 |
| ROC AUC | 0.958 |

Bu değerler, önceki modellerden daha iyi olup özellikle F1 ve AUC açısından zirvededir.

## 13.4 Sınıflandırma Raporu

| **Sınıf** | **Precision** | **Recall** | **F1-score** | **Support** |
| --- | --- | --- | --- | --- |
| 0 – Borcunu ödeyen | 0.92 | 0.99 | 0.95 | 954 |
| 1 – Borcunu ödeyemeyen | 0.96 | 0.67 | 0.79 | 238 |

* Model, riskli müşterilerin %67’sini başarıyla tahmin etmiş ve neredeyse %96 precision sağlamıştır.
* F1 skoru %79’a ulaşarak hem recall hem de precision dengesinde yüksek başarı göstermiştir.

## 13.5 Confusion Matrix

**metin, ekran görüntüsü, diyagram, yazı tipi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

* 157 riskli müşteri doğru tahmin edilmiş
* 81 riskli müşteri gözden kaçmış
* 13 masum müşteri yanlışlıkla riskli olarak sınıflanmış

## 13.6 ROC Eğrisi

**metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

* AUC skoru: **0.958**
* Modelin sınıfları ayırma başarısı oldukça yüksektir

## 13.7 Sonuç

RandomizedSearchCV ile optimize edilmiş XGBoost modeli:

* Tüm modeller içinde **en yüksek F1 ve ROC AUC skorunu** üretmiştir
* Riskli müşterileri tespit etmede güçlü bir **precision-recall dengesi** yakalamıştır
* Uygulama açısından en dengeli ve güvenilir model olarak seçilmiştir

# 14. Tüm Modellerin Karşılaştırması

## 14.1 Performans Karşılaştırma Tablosu

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **ROC AUC** |
| Lojistik Regresyon | 0.737 | 0.391 | 0.567 | 0.463 | 0.761 |
| Random Forest | 0.913 | 0.947 | 0.597 | 0.732 | 0.964 |
| XGBoost | 0.918 | 0.912 | 0.651 | 0.760 | 0.946 |
| XGBoost (RandomSearchCV) | **0.928** | **0.958** | **0.668** | **0.787** | **0.958** |

## 14.2 Değerlendirme ve Yorum

* **Lojistik Regresyon**, temel model olarak referans niteliğinde kalmıştır. Precision düşüktür, fakat yorumlanabilirliği yüksektir.
* **Random Forest**, çok yüksek ROC AUC ve precision değeri ile dikkat çekmektedir. Ancak recall değeri, özellikle riskli müşteri tahmini açısından XGBoost'un gerisindedir.
* **XGBoost (GridSearch)**, dengeli metrikler ile F1 skoru ve recall alanlarında öne çıkmıştır.
* **XGBoost (RandomizedSearchCV)**, tüm metriklerde en başarılı modeli temsil etmektedir. Yüksek F1 skoru ve ROC AUC ile hem dengeyi hem ayrım gücünü sağlamıştır.

# 15. Sonuçlar

Bu projede uygulanan modeller sonucunda aşağıdaki bulgular elde edilmiştir:

* **XGBoost (RandomizedSearchCV)** modeli, kredi riski tahmini açısından en başarılı algoritma olarak öne çıkmıştır.
* Bu model %92.8 doğruluk oranı, %95.8 precision, %66.8 recall ve %78.7 F1 skoru ile tüm modellerin önüne geçmiştir.
* ROC AUC skoru 0.958 ile sınıf ayrımında çok güçlü bir performans sunmuştur

# 16. Gözetimsiz Öğrenme – KMeans Kümeleme

Veri seti üzerinde gözetimsiz öğrenme yaklaşımı uygulanmış ve KMeans algoritması ile benzer özellik gösteren müşteri grupları (kümeler) oluşturulmuştur. Bu işlem, müşteri segmentasyonu, kredi politikası oluşturma veya hedefleme stratejileri açısından değerli içgörüler sağlayabilir.

## 16.1 Uygulanan Adımlar

* **Kategorik eksik veriler** “Unknown” olarak doldurulmuş,
* **Sayısal eksik veriler**, IterativeImputer ile tahmin edilmiştir,
* **Label Encoding** ile kategorik sütunlar sayısallaştırılmıştır,
* BAD (hedef değişken) çıkarılarak yalnızca **bağımsız değişkenler** kullanılmıştır,
* Tüm değişkenler **StandardScaler** ile ölçeklenmiştir.

## 16.2 Elbow Yöntemi ile Optimal Küme Sayısı

**metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

3 küme ile KMeans modeli oluşturulmuştur.

## 16.3 PCA ile 2 Boyutlu Görselleştirme

**metin, diyagram, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

Küme sonuçları, PCA ile 2 bileşene indirgenerek aşağıda görselleştirilmiştir. Grafik, kümelerin birbirinden makul düzeyde ayrıldığını göstermektedir.

## 16.4 Küme Özelliklerinin Karşılaştırması

| **Küme** | **BAD Oranı** | **Ortalama LOAN** | **Ortalama MORTDUE** | **Ortalama VALUE** | **Ortalama DEROG** | **Ortalama DELINQ** | **Ortalama CLAGE** | **Ortalama DEBTINC** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | **0.59** | 17.820 | 60.256 | 88.046 | 1.69 | 2.18 | 149.1 | 36.3 |
| 1 | 0.17 | 15.889 | 53.186 | 79.414 | 0.12 | 0.25 | 165.8 | 32.6 |
| 2 | 0.14 | **26.514** | **124.513** | **169.658** | 0.11 | 0.33 | **227.0** | **28.1** |

**Küme 0**, en yüksek BAD oranına sahip gruptur (%59). Bu grup risklidir: düşük kredi skoru göstergeleri ve yüksek borç oranları (DEBTINC) dikkat çeker.

**Küme 1**, düşük riskli müşteri grubudur. Bu segment potansiyel “güvenli müşterileri” temsil eder.

**Küme 2**, yüksek kredi kullanan ve aynı zamanda düşük BAD oranına sahip güçlü bir müşteri segmentidir. Borcunu büyük ölçüde ödeyebilen ve finansal kapasitesi yüksek bir grubu temsil eder.

KMeans algoritmasıyla veri üzerinde yapılan gözetimsiz kümeleme çalışması, müşteri segmentasyonu açısından anlamlı ve yorumlanabilir gruplar ortaya koymuştur. Bu kümeler, kredi risk puanlaması dışında, daha stratejik analizlerde (örneğin pazarlama, kredi limiti belirleme vb.) kullanılabilir hale gelmiştir.