**Veri Analizi ve Modelleme Raporu**

**1. Veri Setinin Tanımlanması**

**Veri Seti Bilgisi:**

* Veri seti, bankaların finansal oranlarını ve performans ölçütlerini içermektedir.
* Toplamda 41 gözlem ve 9 değişken bulunmaktadır. Seçilen değişkenler şunlardır:
  + (Özkaynak+Kar) / (T.Aktifler+Gayrınakdi Krediler)
  + T.Krediler / T.Aktifler
  + Takipteki Krediler / T.Krediler
  + YP Aktifler / YP Pasifler
  + YP Likit Aktifler / YP Pasifler
  + Net Dönem Karı / Ortalama T.Aktifler
  + Net Dönem Karı / Ortalama Özkaynaklar
  + Net Dönem Karı / Ortalama Ödenmiş Sermaye

**Tanımlayıcı İstatistikler:**

* Veri setindeki sütunlar float türüne dönüştürülmüştür.
* Eksik değer bulunmamaktadır.
* Her bir değişkenin ortalama, standart sapma, minimum ve maksimum değerleri hesaplanmıştır.

1.Çoklu bağlantılılık kontrolü:

Öncelikle Ratio4, Ratio5, Ratio6, Ratio8, Ratio11, Ratio12, Ratio13, Ratio14 sütunları dışındakileri silerek başlıyoruz. Çünkü bizim ilgilendiğimiz sütunlar bunlar.

**Almak istediğimiz sütunları seçiyoruz ve bankalar(string) değişkenlerini de ekliyoruz ve sonrasında analiz yapılmaya hazırlık yapıyoruz.**

**Aşağıdaki soruların adım adım kodlamayla kontrollerini yapıyoruz.**

**-Tüm sütunların sayısal olup olmadığını kontrol edin**

**-Veri kümesinde eksik değerler var mı?**

**-Veri kümenizin boyutlarını kontrol edin.**

**Korelasyon Matrisi:**

* Korelasyon matrisi incelenmiştir ve yüksek korelasyon değerleri tespit edilmiştir.
* Örneğin, (Özkaynak+Kar) / (T.Aktifler+Gayrınakdi Krediler) ile Net Dönem Karı / Ortalama Ödenmiş Sermaye arasında güçlü bir ilişki bulunmuştur (0.92).

**Variance Inflation Factor (VIF):**

* Çoklu doğrusal bağlantı (multicollinearity) VIF kullanılarak analiz edilmiştir.
* VIF değerleri:
  + YP Aktifler / YP Pasifler: 17.04 (Çok yüksek, bağımlılık göstergesi).
  + Takipteki Krediler / T.Krediler: 1.52 (Bağımlılık riski düşük).
  + Net Dönem Karı / Ortalama Özkaynaklar: 1.03 (Bağımsız değişken olarak güvenilir).
* **Sonuç:**
* YP Aktifler / YP Pasifler değişkeni, çoklu doğrusal bağımlılığa sebep olduğu için çıkarılabilir.

Çoklu bağlantılılık (multicollinearity), bağımsız değişkenler arasında yüksek düzeyde ilişki olduğunda ortaya çıkar.

Hangi değişkenleri seçmeliyim?

Bağımsız değişkenlerin (predictors) ilişki durumunu incelemeliyiz. Genellikle bağımsız değişkenlerin birbirine çok yakın korelasyon göstermemesi gerekir. Örnek olarak:

* (Özkaynak+Kar) / (T.Aktifler+Gayrınakdi Krediler)
* T.Krediler / T.Aktifler
* YP Aktifler / YP Pasifler
* Net Dönem Karı / Ortalama T.Aktifler
* Net Dönem Karı / Ortalama Özkaynaklar

**2. Chernoff Yüzü Görselleştirme**

Chernoff yüzleri, veri setinin ilk 10 gözlemi için görselleştirilmiştir. Bu yöntemle değişkenlerin etkisi, yüz ifadeleri ve şekilleriyle anlaşılmıştır.

* **Sonuç:**
  + Takipteki Krediler / T.Krediler gibi yüksek varyansa sahip değişkenler, yüzlerin genel ifadesinde büyük farklılıklar yaratmıştır.
  + Görsel analiz, verilerin etkili bir şekilde özetlenmesini sağlamıştır.

3.Eksik Verili versiyonunu oluşturma

* Eksik verili bir versiyonunu oluşturuyoruz.
  + Eksik verili bir veri setini ortalama ile doldurursak eğer bize sapmalı sonuç vermeyip analizimi etkilemez eğer farklı değerler atarsak küçük veri setleri için problem oluşabilir.
  + En doğru yol ortalama değerlerini sütunların eksik veri yerlerine atamaktır.

Bu grafikten de anlayacağımız üzere Takipteki Krediler / T.Krediler de aykırı değerler mevcut. Sadece

Büyük değerlerde aykırı gözlemler mevcut.Burdan da aşırı derecede bazı bankalarda takipte kredi oranının etkileyen bazı kötü işlemler olduğunu gösterir

Bağımlı değişkenimizde daha fazla uç değer olması ne kadar etkilendiğini bize göstermiş oluyor.

**Eksik Veri Oluşturma:**

* Veri setine rastgele %10 oranında eksik değer eklenmiştir.
* Eksik veriler sütun bazında ortalama ile doldurulmuştur.

**Eksik Verilerin Dağılımı:**

* Eksik veriler sütun bazında rastgele dağıtılmıştır ve MCAR (Missing Completely at Random) varsayımını sağlamıştır.

**Sonuç:**

* Ortalama doldurma yöntemi, veri setindeki sapmayı minimumda tutmuş ve analiz sonuçlarını etkilememiştir.

**4. Aşırı Değer Analizi**

**Boxplot Analizi:**

* Takipteki Krediler / T.Krediler değişkeninde aşırı değerler tespit edilmiştir.
* IQR yöntemi kullanılarak alt ve üst sınırlar belirlenmiştir:
  + Alt sınır: −95.07
  + Üst sınır: 162.69

**Sonuç:**

* Aşırı değerlerin çoğu belirli bankalara özgüdür ve bu durum operasyonel hatalardan kaynaklanabilir.
* Aşırı değerlerin çıkarılması veya dönüşüm yapılması önerilmiştir.

**5. Çoklu Doğrusal Bağlantı Normallik Varsayımı ve Dönüşüm**

Min-max normalizasyon, analizlerin daha tutarlı ve dengeli yapılmasını sağlamıştır.Normalize edildiğinde, tahmin değerleri daha düzenli ve beklenebilir hale gelir. Bu, modelin daha dengeli çalıştığını gösterebilir.

RFE, özellik seçimi için kullanılan bir tekniktir. Özellik seçimi modelin performansını iyileştirmek, aşırı uyumu (overfitting) önlemek ve eğitim süresini kısaltmak amacıyla gereksiz veya anlamlı olmayan özellikleri (değişkenleri) veriden çıkarmak için yapılan bir işlemdir.

**Tek Değişkenli Normallik Analizi:**

* Shapiro-Wilk testi uygulanmıştır:
  + Net Dönem Karı / Ortalama T.Aktifler p-değeri < 0.05; normal dağılmamaktadır.

**Çok Değişkenli Normallik Analizi:**

* Mardia's test sonuçlarına göre veri seti çok değişkenli normal dağılım varsayımını sağlamamaktadır.

**Dönüşüm ve Normalizasyon:**

* Min-max ölçeklendirme uygulanmıştır.
  + Tüm değişkenler 0-1 aralığına getirilmiştir.
  + Dönüşüm sonrası normallik testi tekrarlanmış ve iyileşme görülmüştür.

**Sonuç:**

Min-max normalizasyon, analizlerin daha tutarlı ve dengeli yapılmasını sağlamıştır.Normalize edildiğinde, tahmin değerleri daha düzenli ve beklenebilir hale gelir. Bu, modelin daha dengeli çalıştığını gösterebilir.

RFE, özellik seçimi için kullanılan bir tekniktir. Özellik seçimi modelin performansını iyileştirmek, aşırı uyumu (overfitting) önlemek ve eğitim süresini kısaltmak amacıyla gereksiz veya anlamlı olmayan özellikleri (değişkenleri) veriden çıkarmak için yapılan bir işlemdir.

1. **Eğitim RMSE**: 2.54e-13, modelinizin eğitim verisi üzerinde hatasının çok düşük olduğunu gösteriyor. Bu, modelin eğitim verisiyle çok iyi uyum sağladığını ve hatasının neredeyse sıfır olduğunu ifade eder.
2. **Test RMSE**: 3.57e-13, test verisi üzerindeki hata da oldukça düşük. Model, eğitim setindeki performansını test setine de başarılı bir şekilde taşıyabilmiş gibi görünüyor.
3. **Eğitim MAPE**: 3.36e-12, bu değer de eğitim verisi üzerindeki hata oranının çok düşük olduğunu gösteriyor. Yüzde olarak neredeyse sıfır hata demek, modelin genel eğilimleri doğru bir şekilde yakaladığını ifade eder.
4. **Test MAPE**: 3.37e-12, test verisi üzerinde de benzer bir durum söz konusu, modelin genel hata oranı test setinde de çok düşük. Bu, modelin genelleme kabiliyetinin yüksek olduğunu gösteriyor.

**Genel Değerlendirme**

Bu rapor, bankaların finansal verilerinin analiz edilmesine ve modelleme için uygun hale getirilmesine yönelik işlemleri kapsamaktadır. Elde edilen bulgular:

1. Veri setinde bazı bağımsız değişkenler arasında yüksek korelasyon bulunmaktadır.
2. Eksik verilerin ortalama ile doldurulması analizin doğruluğunu etkilememiştir.
3. Normallik varsayımı dönüşüm teknikleri ile iyileştirilmiştir.
4. Aşırı değerlerin etkisinin azaltılması, modellerin daha güvenilir sonuçlar üretmesine katkı sağlamıştır.