YBS Ansiklopedi
www.YBSAnsiklopedi.com
Cilt 11, Sayı 2, Haziran 2023

Perakende Satışlarda Anomali Tespiti ve Performans Analizi

Anomaly Detection in Retail Sales and Performance Analysis

Gülay GÜRGEN[1]

[1]Kütahya Dumlupınar Üniversitesi, Bilgisayar Mühendisliği, gurgenngulay@gmail.com

ÖZET

Artan veri miktarı sonucunda oluşan sorunlarla birlikte çözümler de şekillenmiştir. Anomali tespiti üretim, sağlık güvenlik, finans hatta uzay çalışmalarında önemli bir yere sahiptir. Gelecekte meydana gelebilecek olumsuz durumlardan örneğin; fabrikalarda iş kazasına yol açabilecek bir makinenin arıza durumunun önceden tespiti, hastalıkların tahmininde sağlık sektöründeki çalışanların iş yükünü hafifletmede, acil durumlarda karar verme süresini kısaltmada, bankacılık sektöründe kayıp/çalıntı gibi durumların tespitinde ve daha birçok problemin çözümünde ya da süreci iyileştirmede anomali tespiti kullanılabilir. Bu makale, perakende satış verilerini kullanarak anomali tespiti yöntemlerini incelemeyi amaçlamaktadır. Anomaliler, perakende sektöründe önemli bir zorluk oluştururken, etkin bir şekilde anormal satışları tespit etmek, işletmelerin gelir kaybını azaltmak ve dolandırıcılık gibi olumsuz etkileri minimize etmek açısından kritik bir öneme sahiptir. Etiketlenmemiş veriler üzerinde hangi tekniklerin, algoritmaların kullanıldığı ve teknikler hakkında ayrıntılı bilgiler verilmiştir. Veri kümesi üzerinde de bu tekniklerden AOM, Autoencoder, AvgkNN, Feature Bagging, GMM, Isolation Forest, kNN, LOF, LSCP, MCD, OCSVM, PCA ve SO-GAAL kullanılmış olup tercih edilme sebepleri ve sonuçları hakkında çıkarımlar ilgili bölümlerde paylaşılmıştır. Sonuç olarak, makale, perakende satış verilerinde anomali tespiti için denetimsiz, yarı denetimli ve diğer yöntemleri karşılaştırmalı bir şekilde değerlendirir. Her bir yöntemin güçlü ve zayıf yönlerini belirleyerek, perakende sektöründeki işletmelerin anormal satışları daha etkin bir şekilde tespit etmelerine yardımcı olmayı hedefler. Anomali tespitinde kullanılacak yöntemin, veri kümesinin özelliklerine ve kullanım senaryosuna göre seçilmesi önemlidir.

Anahtar Kelimeler: Anomali Tespiti, Denetimsiz, Yarı Denetimli

ABSTRACT

As a result of the increasing amount of data, solutions have also been shaped along with the problems. Anomaly detection has an important place in production, health and safety, finance and even space studies. For example, anomaly detection can be used in predetermining the failure of a machine that may cause a work accident in factories, in predicting diseases, in alleviating the workload of employees in the health sector, in shortening the decision-making time in emergencies, in detecting situations such as loss/theft in the banking sector, and in solving many other problems or improving the process. This paper aims to examine anomaly detection methods using retail sales data. While anomalies pose a significant challenge in the retail industry, effectively detecting anomalous sales is critical to reduce revenue loss and minimise negative impacts such as fraud. The techniques and algorithms used on unlabelled data and detailed information about the techniques are given. One of these techniques, AOM, Autoencoder, AvgkNN, Feature Bagging, GMM, Isolation Forest, kNN, LOF, LSCP, MCD, OCSVM, PCA and SO-GAAL, were used on the dataset and the reasons for their preference and conclusions about their results are shared in the relevant section. In conclusion, the paper comparatively evaluates unsupervised, semi-supervised and other

methods for anomaly detection in retail sales data. By identifying the strengths and weaknesses of each method, it aims to help businesses in the retail sector to detect anomalous sales more effectively. It is important to select the method to be used in anomaly detection according to the characteristics of the data set and the usage scenario.

Keywords: Anomaly Detection, Unsupervised, Semi-Unsupervised

1. GİRİŞ

Anomali tespiti, bir sistemin veya olayın normal davranışından sapmaları tanımlayarak, anormal durumları belirleme sürecidir [1]. Bu süreç, çeşitli disiplinlerde bilgisayar biliminde, güvenlikte, sağlık sektöründe ve endüstri gibi çeşitli alanlarda kullanılmaktadır [2]. Günümüzde artan veri miktarı ile oluşan ihtiyaçlarda şekillenmiştir. Büyük miktardaki verinin toplanması ve işlenmesiyle birlikte anormallikleri tespit etmek önem teşkil etmektedir. Örneğin, sağlık verilerinden hastalıkları erken teşhis etmek, finansal işlemlerde dolandırıcılığı belirmek veya bir ağdaki saldırıları tespit etmek gibi birçok uygulama alanı bulunmaktadır.

Anomali tespitinde, kullanılan yöntemler arasında makine öğrenmesi, kümeleme, derin öğrenme ve istatistiksel tabanlı yöntemler gibi farklı yaklaşımlar bulunur [1, 2, 3]. İstatistiksel yöntemler, verinin istatistiksel özelliklerini analiz ederek anormal verileri belirlerken, makine öğrenmesi algoritmaları veri kümesindeki normal örüntüleri öğrenerek anormallikleri tespit eder [4]. Kümeleme algoritmaları, veri noktalarına benzer özelliklere göre gruplara ayırarak anormallikleri tespit ederken, derin öğrenme çok katmanlı yapay sinir ağlarından oluşan bir tekniktir ve karmaşık örüntüleri tanımlayabilir [1, 5]. Anomali tespiti, birçok sektörde yaygın olarak kullanılan ve sürekli geliştirilen bir alandır. Yeni yöntemlerin ve algoritmaların ortaya çıkmasıyla birlikte, anomali tespiti sistemlerinin daha hassas, hızlı ve güvenilir hale gelmesi hedeflenmektedir. Veri güvenliği, sistem bakımı, hata tespiti ve önleme gibi birçok alanda büyük faydalar sağlamaktadır.

Özlem Örnek, Eyyüp Gülbandılar ve Ahmet Yazıcı tarafından gerçekleştirilen bir çalışmada, akıllı fabrikalarda otonom taşıyıcılar için bulanık mantık tabanlı anomali tespiti yapılmıştır. Kullanılan öznitelikler ortalama hız, yaya yoğunluğu, işlem süresi gibi niteliklerken, sonucunda anomali olma oranı ve anomali durumu bulunmuştur. Eşik değerin belirlenmesinde ROC (Receiver Operating Characteristic, Alıcı İşletim Karakteristiği) analizi kullanılmıştır. Yapılan çalışmada 2005 adet veride %71,87 doğruluk oranı ve 0,38 yanlış negatif hata oranı elde edilmiştir [6].

2014 yılında yayınlanan "Tucker3 Ayrıştırması Kullanarak Olay ve Anomali Tespiti" adlı makalede telekomünikasyon ağlarında arıza tespiti için etiket verisi gerektirmediğinden dolayı denetimsiz öğrenme yöntemi üzerinde durmuşlardır. Ayrıca ağ verilerinin zamansal sıralı yapıda olması da bu kararı desteklemiştir. Çalışmanın sonucunda Tucker3 ayrışımının IP/TV ağındaki kullanıcılar için güvenilir bir yöntem olduğuna kanaat getirmişlerdir [3].

Anomali tespiti uzay araştırmalarında da kullanılmıştır. "Uzay Aracı Anomali Tespit Problemine Bir Yaklaşım Kernel Özellik Uzayını Kullanma" adlı makale de uzay araçlarında anomali tespiti üzerine yeni bir yöntem önerilmiştir. Uzay aracı için Schölkopf, Smola ve Muller tarafından önerilen kernel PCA (Principle Component Analysis, Temel Bileşen Analizi) yöntemini uygulamışlardır [7, 8].

Sıralı veri analizi, gelecekteki trendleri tahmin etmek, anormallikleri tespit etmek, desenleri ortaya çıkarmak (Pattern Mining, Örüntü Madenciliği), sınıflandırmak veya kümelemek gibi bir dizi hedefe yönelik kullanılır. "Anomali Tespiti İçin Sıralı Verilerden Öğrenme" adlı çalışmada sıralı verilerde anomali tespiti üzerinde durulmuştur ve ayrık diziler kullanılmıştır. Yapılan başka bir çalışmada çeşitli veri kaynakları ile anomali ve aykırı değer tespit yöntemlerinin ayrıntılı bir incelemesi sunulmuş ve çoklu uygulama alanlarında anomali tespit tekniklerini altı grup altında toplamışlardır: sınıflandırma temelli, kümeleme temelli, yapay sinir ağı (YSA) temelli, istatistiksel, spektral ve bilgi teorik yöntemlerdir [4, 9]. Yolaçan tarafından yapılan çalışma sonucunda sıralı veriler üzerinde anomali tespitini iyileştirmek için yeni bir yaklaşım sunmuştur. Ayrık, sembolik ve sıralı veri kümesinde anomali tespiti teknikleri üzerinde durmuştur [10].

Erhan Yılmaz'ın yapmış olduğu "Makine Öğrenmesi Tabanlı Kullanıcı Davranış Analizi ile Bilgisayar Sistemlerine Giriş Kayıtlarında Anomali Testi" adlı yüksek lisans tezinde, kNN (k-Nearest Neighbors, k-En Yakın Komşu), LOF (Local Outlier Factor, Yerel Aykırı Değer Faktörü), COF (Connectivity-based Outlier Factor, Bağlantı Tabanlı Aykırı Değer Faktörü), LoOP (Local Outlier Probability, Yerel Aykırı Değer Olasılığı), INFLO (Improving Influenced Outlierness, Etkilenen Aykırılığın Geliştirilmesi), CBLOF (Clustering-Based Local Outlier Factor, Kümeleme Tabanlı Yerel Aykırı Değer Faktörü) ve LDCOF (Local Density Cluster-Based Outlier Factor, Yerel Yoğunluk Küme Tabanlı Aykırı Değer Faktörü) algoritmalarının gösterdikleri performansı karşılaştırmıştır. Yapılan çalışma sonucunda, yakınlık tabanlı algoritmaların, kümeleme tabanlı algoritmalara göre anomalileri tespit etmede daha iyi performans gösterdikleri belirtilmiştir. AUC (Area Under the Curve, Eğrinin Altındaki Alan) değerlerinin ortalaması sonucunda, kullanılan veri kümesi için en iyi performansı yakınlık tabanlı anomali tespiti algoritmalarından INFLO göstermiş olup algoritmaların hesaplama süreleri karşılaştırıldığında ise kümeleme tabanlı algoritmaların, yakınlık tabanlı algoritmalardan daha hızlı sonuç verdiği belirtilmiştir [11, 12].

Yapılan literatür taraması ile anomali tespitinin her disiplinde var olduğu görülmektedir. Kullanılan veri kümesine ve ihtiyaçlara göre seçilebilecek algoritma çeşitlilik gösterdiği gibi baz alınan algoritmanın farklı versiyonları da literatürde bulunmaktadır. Giriş bölümünde daha geleneksel algoritmalar yer alırken bu algoritmaların türevlerinden, geliştirilen algoritmalar hakkında araştırmalar yöntemler bölümünde yer almaktadır. Bu makalede, anomali tespitinde kullanılan yöntemlerin bazıları ele alınacak olup ilerideki çalışmalar için fikir verme amacı taşımaktadır. Denetimsiz anomali tespit teknikleri, yarı denetimli ve yapay sinir ağı temelli teknikler üzerinde durulacaktır. Gerçek aykırı değerler bilinmediği durumlarda, anomali kaynaklarının oluşum sebepleri ya da kimin sebep olduğu gibi yorumlamaların yapılabilmesi adına fikir verme amacı taşımaktadır. Bu amaç doğrultusunda, literatürdeki çalışmalar taranmıştır.

2. METODOLOJİ

Bu bölümde veri kümesi ve kullanılan algoritmalar hakkında ayrıntılı bilgiler yer almaktadır. Anlatımda kullanılan tablo ve grafikler ile okuyucunun problemi kavramasına yardımcı olma amacı taşımaktadır. Veri ön işleme, özellik mühendisliği, model kurulumu ve sonuçlar hakkında özet bilgiler de bulunmaktadır.

2.1 Veri Kümesi

Bu çalışmada kullanılan veri kümesi, şube ve bayilerden toplu satış perakende verileridir. Satış verileriyle ilgili temel bilgileri içerir. Firmanın satış acenteleri/bayileri ve şubeleri bulunmakta olup veri dosyasında müşteri alanında sadece şube/bayi bilgileri yer almaktadır. Veri kümesi toplam 29103 kayıttan oluşmaktadır. Veri kümesindeki öznitelikler, işlem numarası, müşteri numarası, ürün kodu, satış fiyatı, satış adedi, indirim tutarı ve tarih bilgisi olmak üzere 7 adet öznitelik bulunmaktadır. Yapılan özellik mühendisliği adımlarından sonra öznitelik sayısı 23'e çıkarılmıştır. Kullanılan veri kümesinin almış olduğu değerler ve veri tipleri ile ilgili özet bilgiler Tablo 1'de sunulmuştur. Özelliklerle ilgili tanımlayıcı aralıkların belirlenmesi için kullanılan kaynaklar belirtilmiştir.

Tablo 1: Çalışmada kullanılan satış perakende veri kümesinin bilgilerini içermektedir.

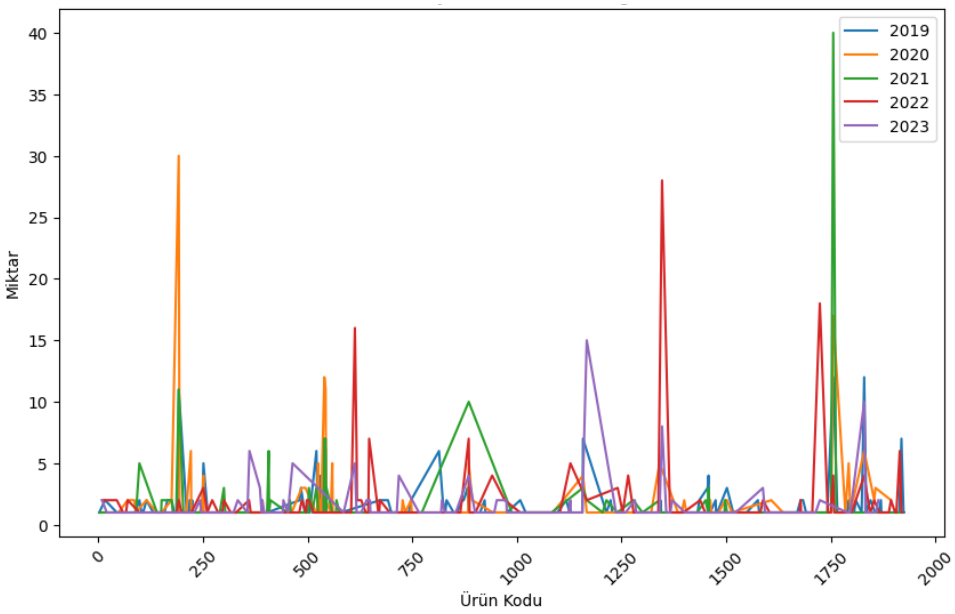
Özellikler	Değerler	Açıklama
İşlem Numarası	Kategorik	Her tek seferlik işlem için kesilen faturaya atanmış kimlik numarasıdır.
Müşteri Numarası	Kategorik	Her müşteriye özel olarak atanmış kimlik numarasıdır.
Ürün Kodu	Kategorik	Her ürün için özel olarak atanmış ürün kimlik numarasıdır.

Özellikler	Değerler	Açıklama
Toplam Satış Fiyatı	Nümerik	Müşterinin tek seferde yapılan işlem için ödediği genel toplamdır.
Satış Adedi	Nümerik	İşlem özelinde müşterinin üründen kaç adet aldığı bilgisini vermektedir.
İndirim Tutarı	Nümerik	İşlem için uygulanmış indirim tutarı bilgisini vermektedir.
Tarih	Kategorik	Satın almanın gerçekleştiği zaman bilgisini verir.
Birim Satış Fiyatı	Nümerik	Ürünün birim fiyatını verir.
İndirim Kontrol	Kategorik	Müşteriye indirim uygulanmış mı kontrol eder ve sonuca göre evet ya da hayır cevabı verir.
Segment	Kategorik	Her müşterinin tüm zamanlara göre gelir getirisini hesaplar ve diğer müşterilere göre kıyaslayıp üç grup altında sınıflandırır.
Satın Alma Frekansı	Kategorik	Her müşterinin tüm zamanlara göre satın aldığı ürün adedine göre alışveriş sıklığını verir.
Satın Alma Segmenti	Kategorik	Tüm müşterilerin satın alma geçmişlerini karşılaştırıp gruplar ve üç grup altında sınıflandırır.
Adede göre en az alım	Kategorik	Toplam satış adedine göre en az alım yapan müşteri bilgisini verir.
Adede göre en çok alım	Kategorik	Toplam satış adedine göre en çok alım yapan müşteri kimliğinin bilgisini verir.
Müşteri Ürün Adet	Nümerik	Her müşterinin hangi ürünü kaç adet aldığını hesaplar.
Müşterilerin Popüler Ürünü	Kategorik	Her müşteri için en fazla tercih edilen ürünün bilgisini verir.
Müşteriler Arasında Popüler Olmayan Ürün	Kategorik	Her müşteri için en az tercih edilen ürünün bilgisini verir.
Yıllık En Tercih Edilen Ürün	Kategorik	Yıllara göre müşterilerin en çok tercih ettiği ürünlerin bilgisini verir.
Yıllık En Tercih Edilen Ürün Sayısı	Nümerik	Her yıl için müşterilerin en çok tercih ettiği ürünlerin sayısını tutar.
Yıllık En Az Tercih Edilen Ürün Sayısı	Nümerik	Her yıl için müşterilerin en az tercih ettiği ürünlerin sayısını tutar.
Toplam Satışta En Fazla Alım Yapan Müşteri	Kategorik	Toplam satış ücretine göre en fazla alım yapan müşteri.
Toplam Satışta En Az Alım Yapan Müşteri	Kategorik	Toplam satış ücretine göre en az alım yapan müşteri.
Ay	Kategorik	Tarih bilgisinin aylara bölünmesinden oluşmaktadır.
Yıl	Kategorik	Tarih bilgisinin yıllara bölünmesinden oluşmaktadır.

İlk aşamada keşifçi veri analizi ile veri kümesinin ön değerlendirilmesi yapılmıştır. Satış dağılım grafikleri ve müşteri hareketleri incelendikten sonra veri kümesinde zenginleştirilmeye gidilmiştir. Toplam satış gelir bilgisinin satın alma adedine bölünmesiyle birim satış geliri hesaplanmıştır. İndirim uygulanmış işlemler tespit edilip, ileride anomali tespit aşamasında yorum yapabilmek adına yeni bir değişkende tutulmuştur. Toplam satış gelirinin tüm veri kümesi üzerinde ortalaması alınıp, çıkan sonuç yapılan her fatura işleminin genele oranla düşük, ortalama ve yüksek gelir getirisi olmak üzere üç kategoriye ayırır, segment adlı değişkende bilgileri tutulur. Satın alma frekansı hesaplanırken satış adedi yerine işlem numarası dikkate alınmıştır. İşlem bilgilerinin kullanılmasının sebebi, her bir faturanın benzersiz bir kimlik numarası olmasıdır. Her bir fatura, müşterinin gerçekleştirdiği ayrı bir satın alma işlemini temsil eder. Dolayısıyla, müşterilerin toplam satın alma sayısını hesaplamak için her bir faturanın sayısını kullanmak uygun bir yaklaşımdır. Eğer satın alma adedi değişkenindeki değerler kullanılmış olsaydı, satın alma frekansı hesaplanırken müşterinin toplam ürün miktarını dikkate alırdı. Bu, müşterinin her bir satın alma işleminde kaç adet ürün satın aldığını gösterir. Ancak, bu durumda aynı ürünü birden fazla kez satın alan müşterilerin satın alma frekansı yüksek görünebilir. Bu nedenle, satın alma frekansı

hesaplanırken işlem numarası dikkate alınmıştır. Satın alım frekansları sayısal bir değişkende tutulurken, kategorik olarak da düşük, ortalama ve yüksek frekans olmak üzere üç gruba ayrılmıştır.

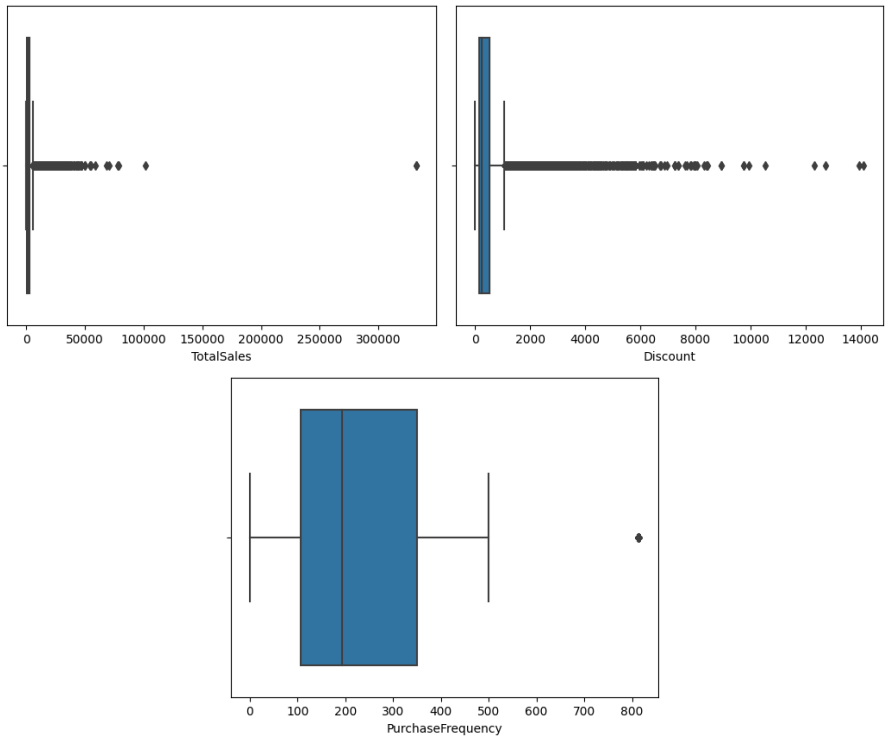
Keşifçi veri analizi aşamasında üretilmiş olan değişkenlerin analizi yapılmıştır. Bu analiz sonucunda, her müşterinin hangi ürünlerden kaç adet aldığı hesaplanıp yeni bir değişkende tutulmuştur. Bu işlem sonrasında, her müşterinin en fazla ve en az tercih ettiği ürünler bulunmuştur. Yıl bazında müşterilerin almış olduğu ürünler ve sayıları birlikte Figür 1'de verilmiştir.



Figür 1: Yıllara göre müşterilerin tercih ettiği ürünler.

Her müşteri özelinde toplam satış adedi ve satış gelirleri hesaplanmıştır. Bu işlemin sonucunda, toplam satış adedine göre en yüksek ve en düşük alım yapan müşteriye ulaşılmıştır. Benzer bir işlemle de toplam satış geliri hesaplamaya katılarak en yüksek ve en düşük para getirisine sahip müşteriler tespit edilmiştir.

Aykırı değer kontrolu yapılırken toplam satış fiyatları (2578), indirim tutarı (2674) ve satın alma frekansında (4153) aykırı değer olduğu tespit edilmiştir. Satın alma frekansı göz ardı edilebilir. Müşterilerin hem satın alma senaryoları birden fazla metriğe bağlı olduğu hem de bu değişken sonradan üretildiği için üretiminden kullanılan değişkenlerden kaynaklı bir aykırı değerlere de sahip olabilir. Bu aşamada, aykırı değerlerin olduğu gibi bırakılmasına karar verilmiştir. Anomali tespitinde faydalı olabileceği düşünülerek aykırı değerlere dokunulmamıştır. Boş ya da geçersiz değerler de kontrol edilmiştir.



Figür 2: Toplam satış, indirim tutarı ve satın alma frekansı aykırı değerlerinin grafikleri paylaşılmıştır.

Yapılan özellik mühendisliği çalışmaların son aşaması olan kategorik değişken dönüşümü yapılmıştır. Label ve One-Hot Encoding işlemleri birlikte kullanılmıştır.

2.2 Yöntemler

Literatürde kullanılan yöntemler taranmış olup veri kümesine en uygun olanlar seçilmiştir. Çalışmada 19 algoritma denenmiş olup 13 algoritma değerlendirmeye alınmıştır. Veri kümesi üzerinde sırasıyla ABOD (Angle-Based Outlier Detection, Açı Tabanlı Aykırı Değer Tespiti), AOM (Average of Maximum, Maksimum Ortalama), Autoencoder (Otomatik Kodlayıcı), AvgkNN (Average k-Nearest Neighbors, Ortalama k-En Yakın Komşu), CBLOF, Feature Bagging (Özellik Torbalama), GMM (Gaussian Mixture Model, Gauss Karışım Modeli), HBOS (Histogram-Based Outlier Score, Histogram Tabanlı Aykırı Değer Puanı), Isolation Forest (İzolasyon Ormanı), kNN, LOCI (Local Correlation Integral, Yerel Korelasyon İntegrali), LOF, LSCP (Locally Selective Combination in Parallel Outlier Ensembles, Paralel Aykırı Değer Topluluklarında Yerel Seçici Kombinasyon), LSTM (Long Short-Term Memory, Uzun Kısa Süreli Bellek), MCD (Minimum Covariance Determinant, Minimum Kovaryans Belirleyici), OCSVM (One-Class Support Vector Machine, Tek Sınıflı Destek Vektör Makinesi), PCA, SO-GAAL (Single-Objective Generative Adversarial Active Learning, Tek Amaçlı Üretken Çekişmeli Aktif Öğrenme) ve SOS (Stochastic Outlier Selection, Stokastik Aykırı Değer Seçimi) algoritmaları kullanılmıştır. Veriler öncelikle bir veri ön işleme aşamasından geçirilmiştir. Bu aşamada, yeni değişken üretimi, istatistiksel analiz, aykırı değer analizi, encoding işlemleri ve veri ölçeklendirilmesi yapılmıştır. ABOD, CBLOF, HBOS, LOCI ve SOS algoritmaları Numba kütüphanesinin çalışmasına yönelik hatalar verdiği için bu algoritmalar çalışma dışında

bırakılmıştır. PyOD (Python Outlier Detection, Python Aykırı Değer Tespiti) ve Numba sürüm farkı uyuşmazlıkları, ortam değişkenlerinin eksik ya da hatalı kurulumu gibi nedenlerin sebep olabileceği düşünülmüş olsa da birçok yöntem denenmiş olup hata giderilememiştir. Bu yüzden de çalışma da kullanılmamak üzere çıkarılmıştır. LSTM algoritmasının çıkarılma nedeni de ilgili bölümde aktarılmıştır. Algoritmalar özelinde çıkarılan tabloların analizleri bulgular ve tartışma bölümünde irdelenmiştir.

2.2.1 Denetimsiz anomali tespiti

Denetimsiz yöntemler, etiketlenmemiş verileri kullanarak anormal örüntüleri tespit etmeye odaklanır [4]. Aykırı değer saptama (Outlier Detection), yoğunluk tabanlı ve kümeleme tabanlı anomali tespiti gibi denetimsiz teknikler, perakende verilerinde anormal satışları tespit etmek için kullanılan yaygın yöntemlerdendir. Denetimsiz anomali tespitinde kullanılan tekniklerde, anormal veri noktaları normal veri noktalarına göre daha az görülür [13]. Bu hipoteze göre, daha az rastlanan veri noktalarını anomalilik olarak nitelendirilmesini sağlar. Denetimli anomali tespit tekniklerinde her veri noktasına etiket atanırken, denetimsiz tekniklerde ise veri noktalarına sayı atanır [14]. Atanan bu sayılar, veri noktasının anomali olma olasılığı hakkında yorum yapmamıza katkı sağlar.
Kullanılan Denetimsiz Anomali Tespit Algoritmaları:

İstatistiksel Teknikler

ABOD

Bir noktanın fark vektörleri arasındaki açıları kullanarak aykırı değerleri tespit etmeyi hedefler, bu sayede boyutsallık sorunun etkileri azaltılmış olur ve parametresiz bir algoritma olması da avantaj sağlar [15]. Hans-Peter Kriegel, Matthias Schubert, Arthur Zimek tarafından yapılan deneylerde, FastABOD, ABOD ve LOF algoritmalarının çoklu veri tabanı boyutlarında performansları karşılaştırılmıştır. ABOD'un diğer yöntemlerle karşılaştırıldığında özellikle de yüksek boyutlu verilerde iyi performans gösterdiği sonucuna varılmıştır. Aynı zamanda bu yöntem, veri madenciliğinde "boyutluluk laneti"nin azaltılmasına yardımcı olduğu belirtilmektedir [15].

HBOS

Her bir özellik için tek değişkenli bir histogram oluşturulur. Kategorik veriler için değerlerin sayıldığı ve göreceli frekansın (histogramın yüksekliği) hesaplandığı bir yöntem kullanılmaktadır, sayısal veriler için ise statik kutu histogramları veya dinamik kutu histogramları kullanılabilmektedir [16]. Çok değişkenli anomali tespitinde, aykırı değer skorları her histogram için ayrı ayrı hesaplanılır ve daha sonra birleştirilir [17]. Markus Goldstein ve Andreas Dengel tarafından yapılan çalışmada, HBOS'un kümeleme tabanlı algoritmalardan 5, yakınlık tabanlı algoritmalardan ise 7 kata kadar daha hızlı olduğunu tespit etmişlerdir. Küresel anomali tespit problemlerini çözebilirken, yerel aykırı değerlerinde bu başarıya ulaşamamıştır [16].

PCA

PCA, veri boyutunu azaltma ve veri kümesinin varyansının ayarlanmasında kullanılan boyut indirgeme tekniğidir [18]. Azaltılan boyuttaki veri uzayında yer alan normal veriler birbirine daha yakın dururken aykırı değerler gruplardan ayrı konumlanmış olurlar. Literatürde PCA'nın zamansal veri noktalarının korelasyonunu dikkate almadığı için düşük performans gösterebileceğini, bu sorunu minimize etmek için de KL (Karhunen-Loeve) genişlemesinin kullanılması önerilmiştir [19].

Tablo 2: PCA Anomali Raporu

Anomali Sayısı	İşlem numarası ve tekrar adedi	Müşteri numarası ve tekrar adedi	Ürün numarası ve tekrar adedi	Genel İndirim Uygulanma Oranı	Genel Maximum Segment Oranı	Genel Maximum Satın Alma Frekansı Oranı
522	9270 (4) 7428 (4) 12884 (3)	420 (39) 375 (37) 57 (37)	540 (44) 192 (37) 543 (35)	%99,04	%78,92 Yüksek Satış	%90,42 Yüksek Frekans

MCD

Veri noktalarının kovaryans matrisinin, determinantı alınarak veri noktalarının dağılımı incelenir. Yüksek determinant değeri normal veriyi temsil ederken düşük determinant ise aykırı değerleri temsil eder [20, 21]. MCD aykırı değerlere karşı dirençli olmasıyla güçlü bir algoritmadır [22].

Tablo 3: MCD Anomali Raporu

Anomali Sayısı	İşlem numarası ve tekrar adedi	Müşteri numarası ve tekrar adedi	Ürün numarası ve tekrar adedi	Genel İndirim Uygulanma Oranı	Genel Maximum Segment Oranı	Genel Maximum Satın Alma Frekansı Oranı
583	12030 (8) 11392 (5) 11125 (4)	57 (44) 375 (38) 52 (34)	1347 (85) 1830 (25) 885 (24)	%94,16	%95,54 Yüksek Satış	%65,18 Yüksek Frekans

Kümeleme Tabanlı Anomali Tespiti

CBLOF

CBLOF algoritmasını geliştiren uzmanlar, tüm veri kümesine odaklanmak yerine yerel veri kümelerinin davranışlarını incelemeyi amaçlamışlardır. Veriyi gruplara ayırmak için 2002 yılında geliştirilen kümeleme algoritması Squeezer' i [23] kullanmışlardır. Squeezer ile küme merkezleri belirlendikten sonra aykırı değer faktörü aşaması başlatılır, veri noktalarının konumları ve birbiriyle olan ilişkileri incelenir ve yapılan hesaplamalar sonucunda CBLOF algoritması aykırı değerleri belirler. Aynı zamanda yayınlanan çalışmada, FindCBLOF adlı algoritma tanıtılmış olup CBLOF tarafından bulunan aykırı değerlerin analizi ve raporlanması için geliştirilmiştir [24].

Şüpheli işlemlerin tespiti için yapılan başka bir çalışmada, CBLOF algoritmasının başarılı yönlerinden birinin veri eksikliğinden etkilenmemesi olduğu belirtilmiştir. Performansının yeni dolandırıcılık yöntemlerinde de başarılı olabileceği ve eğitim verisine ihtiyaç duymadığı için de kısıtlı veri kümelerinde bile başarılı bir performans gösterdiği belirtilmiştir. Verilerin kümelenmesinde bir önceki atıfta bulunulan çalışmaya göre farklı olarak LOF algoritmasını tercih etmişlerdir. Genel aykırı değerlerin tespitindense, daha etkili sonuçlar elde etme amacıyla yerel veri noktalarına odaklanmak istemeleri LOF algoritmasını kullanmalarına sebep olmuştur [25].

GMM

Xingwei Yang, Longin Jan Latecki, Dragoljub Pokrajac yapmış olduğu çalışmada, anomali tespitinde kullanabilmek üzere EM (Expectation-Maximization, Beklenti Maksimizasyonu) yönteminin genel bir optimal versiyonuna bağlı bir yaklaşım sunmuşlardır. Veri kümesinin normal davranışından sapma gösteren noktalarını tespit etmeye odaklanır. Gauss

çekirdeğinin standard sapma değerini parametre olarak kullanır. LOF ve COF gibi algoritmalarının kullanıldığı farklı çalışmalarla karşılaştırıldığında performans olarak daha iyi sonuç verdiği ancak, karmaşıklığının yüksek olduğunu tespit etmişlerdir [26].

Tablo 4: GMM Anomali Raporu

Anomali Sayısı	İşlem numarası ve tekrar adedi	Müşteri numarası ve tekrar adedi	Ürün numarası ve tekrar adedi	Genel İndirim Uygulanma Oranı	Genel Maximum Segment Oranı	Genel Maximum Satın Alma Frekansı Oranı
	11392 (9)	230 (115)	192 (72)			
1456	11123 (6)	375 (78)	1830 (71)	%83,37	%64,14 Yüksek Satış	%82,48 Yüksek Frekans
	12030 (6)	420 (76)	540 (69)			

Yakınlık Tabanlı Anomali Tespiti

SOS

SOS algoritması, bir özellik matrisi veya benzerlik matrisi girdisi alır ve her veri noktası için bir anomali olasılığı çıkarır [27]. Janssens, J.H.M tarafından yapılan çalışmada gerçek ve sentetik veriler olmak üzere 25 veri kümesi kullanılmıştır. SOS ile KNNDD (K-Nearest Neighbour Data Description, K-En Yakın Komşu Veri Tanımı), LOF, LOCI ve LSOD (Least Squares Outlier Detection, En Küçük Kareler Aykırı Değer Tespiti) algoritmalarının performansları aykırı değer skoru grafikleri kullanılarak karşılaştırılmıştır. Nemenyi testi sonucunda, SOS' un performansı diğer dört algoritmaya göre yüksek olduğu ortaya çıkmıştır. Veri pertürbasyonlarına (veri bozulması, gizlenmesi) ve değişen yoğunluklara diğer dört algoritmaya göre daha dayanaklı olduğu ortaya çıkmıştır [28].

kNN

kNN algoritması, belirlenen k değeri kadar komşuluktaki veri noktalarının dağılımını ve yoğunluğunu inceler. Tanımlanan eşik değere göre veri noktalarının birbirine yakın olduğu bölgelerden uzakta olanları aykırı değerler olarak işaretler. Literatürde, kNN algoritması temel alınarak farklı yöntemlerle birlikte aykırı değerlerin tespitinde kullanılmıştır.

2008 yılında yapılan çalışmada, squeezer algoritması kullanılarak veri noktaları kümelere ayrıldıktan sonra kNN algoritması kullanılmıştır. Gereksiz hesaplamalardan kurtulmak için de budama işlemi R-Ağacı ile algoritmaya entegre edilmiştir. Yapılan çalışma yüksek boyutlu, sentetik ve gerçek veri kümelerinde uygulandığında geleneksel kNN'e göre daha iyi performans göstermiştir [29].

Nesnelerin interneti, anlık veri akışı alanında yapılan çalışmada kNN tabanlı bir anomali tespiti yöntemi önerilmiştir. GAAOD (Grid-Based Approximate Average Outlier Detection, Izgara Tabanlı Yaklaşık Ortalama Aykırı Değer Tespiti) algoritmasını üç aşamada ele almak gerekirse, ilk olarak mesafe dağılımı ve komşulukları öğrenildikten sonra bu bilgileri bir arada tutabilmek için veriler ilgili kümelere bölünerek "grid-based index" adlı indekste tutulur. GAAOD algoritması öz yapılandırma özelliği sayesinde, veri noktalarının arasındaki mesafelere uygun ızgara yapısını hücre boyutuna göre ayarlayabilmektedir. Son olarak da aykırı değer adayı veri noktalarının belirlenmesi normal olarak nitelendirilebilecek verilerin de kümeden çıkarılabilmesi için "k-skybands" algoritması temelli bir yaklaşım kullanılmıştır. Bu yapılan çalışma anlık gelişen olayların önceden tahmin edilmesi ya da erken bildirim için çeşitli alanlarda kullanılabilir [30].

Tablo 5: kNN Anomali Raporu

Anomali Sayısı	İşlem numarası ve tekrar adedi	Müşteri numarası ve tekrar adedi	Ürün numarası ve tekrar adedi	Genel İndirim Uygulanma Oranı	Genel Maximum Segment Oranı	Genel Maximum Satın Alma Frekansı Oranı
1456	7042 (4) 13197 (4) 7843 (4)	57 (69) 420 (59) 15 (50)	1756 (89) 540 (70) 192 (68)	%88,73	%61,19 Yüksek Satış	%65,10 Yüksek Frekans

AvgkNN

AvgkNN, kNN ile aynı adımlardan oluşmaktadır. Farklı olarak yapılan işlem ise her bir veri noktasının komşularına uzaklığının ortalaması alınır ve bu ortalama değer, anomali skoru olarak kullanılır.

Tablo 6: AvgkNN Anomali Raporu

Anomali Sayısı	İşlem numarası ve tekrar adedi	Müşteri numarası ve tekrar adedi	Ürün numarası ve tekrar adedi	Genel İndirim Uygulanma Oranı	Genel Maximum Segment Oranı	Genel Maximum Satın Alma Frekansı Oranı
1456	7861 (25) 8405 (19) 7804 (17)	230 (1186) 52 (28) 181 (28)	192 (261) 1756 (97) 221 (94)	%99,45	%8,17 Ortalama Satış	%98,35 Yüksek Frekans

LSCP

LSCP algoritması, diğer anomali tespitinde kullanılan algoritmaları paralel olarak çalıştırarak, en iyi performans gösterenleri seçer ve kullanılan algoritmaların temel zayıflıklarından oluşabilecek sorunları çevresel ilişkilerine dikkat ederek ortadan kaldırmaya çalışır. Her bir veri noktasına özel, yerel bölgeyi belirler ve bu bölgede en iyi performansı gösterebilecek algoritmayı seçer. Son aşamada ise ortaya çıkan sonuçlar birleştirilir ve aykırı değerler tespit edilir. LSCP algoritmasını geliştirenler aynı zamanda farklı varyanslarını da denemiş olup en başarılısını 20 veri kümesi üzerinde test ederek LSCP_AOM algoritması olduğuna kanaat getirmişlerdir [31].

Tablo 7: LSCP Anomali Raporu

Anomali Sayısı	İşlem numarası ve tekrar adedi	Müşteri numarası ve tekrar adedi	Ürün numarası ve tekrar adedi	Genel İndirim Uygulanma Oranı	Genel Maximum Segment Oranı	Genel Maximum Satın Alma Frekansı Oranı
15	348 (2) 359 (2) 332 (1)	230 (3) 250 (2) 267 (1)	525 (2) 1468 (1) 493 (1)	%100	%39,99 Ortalama Satış	%79,99 Yüksek Frekans

LOF

Veri noktalarının yerel komşuluklarına göre yoğunluk miktarının oranlanmasıyla LOF değeri oluşturulur. İncelenen veri noktasının kalabalık bir küme içerisinde konumlanmışken yoğunluğunun yüksek olması beklenir fakat buna rağmen veri noktasının yoğunluğunun düşük olması küme içerisinde kendisini izole ettiğini, yani oraya ait olmadığını ifade eder

böylece LOF değeri yüksek hesaplanır. Tam tersi bir durumda gelişebilir. Çevresinde yer alan veri noktalarının yoğunluğu düşükken, baz alınan veri noktasının yoğunluğunun yüksek olması da bir anomali olduğunun göstergesi olup LOF değerinin yüksek çıkmasına sebep olur [32].

LOF değeri 1'e yakın olduğunda yanlış yorumlamalara sebep olabiliyor. Bu soruna çözüm olarak uzmanlar LDBSCA (Local Density-Based Spatial Clustering of Applications with Noise, Uygulamaların Yerel Yoğunluk Tabanlı Gürültü ile Mekansal Kümelenmesi) algoritmasını kullanımını önermişlerdir. Böylece LOF değeri sadece veri noktasının komşuluklar arasındaki yoğunluk ilişkisinin yanı sıra ait olduğu küme içerisinde de hesaplama yapılır. Çeşitli veri kümelerinde yapılan testler ile de LDBSCA'nın performansının yüksek olduğunu göstermişlerdir [33].

Tablo 8: LOF Anomali Raporu

Anomali Sayısı	İşlem numarası ve tekrar adedi	Müşteri numarası ve tekrar adedi	Ürün numarası ve tekrar adedi	Genel İndirim Uygulanma Oranı	Genel Maximum Segment Oranı	Genel Maximum Satın Alma Frekansı Oranı
	6221 (16)	230 (784)	192 (55)			
2608	7804 (15)	420 (99)	1830 (41)	%92,82	%34,93	%84,43
					Ortalama Satış	Yüksek Frekans
	6222 (14)	424 (69)	1756 (36)			

LOCI

LOCI algoritmasını geliştiren uzmanlar, diğer algoritmalara göre daha hızlı çalışabilen, mikro kümeleri tespit edebilen ve yerel yoğunluğun yanı sıra çoklu küçük küme boyutlarında performans gösterebilen bir yöntem ele almışlardır. Ayrıca her bir veri noktasına özel üretilen grafikler, veri kümesinin davranışını görselleştirmesiyle ve aykırı değerin tespitinde kullanılan eşik değerini kullanıcı yerine algoritmanın belirlemesiyle diğer algoritmalardan farklı olduğunu belirtmişlerdir [34].

2.2.2 Yarı Denetimli Anomali Tespiti

Yarı denetimli öğrenme, etiketlenmemiş verinin daha fazla olduğu durumlarda faydalıdır. Bu durum, veri noktalarını elde etmenin kolay ya da ucuz olduğu durumlarda ortaya çıkar, fakat verilere etiket eklemek zaman, iş yükü ve yüksek maliyetlere neden olur [35]. Yarı denetimli anomali tespiti, etiketli ve etiketsiz verileri birleştirerek daha iyi sonuçlar elde etmeyi amaçlayan önemli bir alandır. Model normal veri noktalarından oluşan eğitim seti üzerinde çalışır, buradan öğrendikleri ile de bu davranışların dışına çıkan veri noktalarını aykırı değer olarak işaretler [12]. Yarı denetimli SVM (Support Vector Machine, Destek Vektör Makinesi), etiketlenmiş verileri sınıflandırma için kullanırken, etiketsiz verileri de anomalileri tespit etmek için kullanabilir Aynı şekilde, GMM tabanlı yöntemler de etiketlenmiş verilerdeki normal örüntüyü öğrenir ve etiketsiz verilerde anormal satışları tespit eder.
Kullanılan Yarı Denetimli Anomali Tespit Algoritmaları:

AOM

Veri kümesi üzerinde AOM algoritması, aykırı değer olmadığı sonucuna varmıştır. Farklı k değerleri ve eşik değerleri kullanılarak yapılan denemelerde de sonuç değişmemiştir.

Tablo 9: AOM Anomali Raporu

Anomali Sayısı	İşlem numarası ve tekrar adedi	Müşteri numarası ve tekrar adedi	Ürün numarası ve tekrar adedi	Genel İndirim Uygulanma Oranı	Genel Maximum Segment Oranı	Genel Maximum Satın Alma Frekansı Oranı
0	-	-	-	-	-	-

2.2.3 Derin Öğrenmeye Dayalı Anomali Tespit Algoritmaları

Autoencoder, LSTM ve SO-GAAL algoritmalarında Keras kütüphanesinden faydalanılmıştır.
Kullanılan Derin Öğrenmeye Dayalı Anomali Tespit Algoritmaları:

Autoencoder

Autoencoder algoritması için aktivasyon fonksiyonu olarak ReLU (Rectified Linear Unit, Yenilenmiş Doğrusal Birim) kullanılmıştır. Algoritma, çıktı katmanının girdi katmanına göre daha az kayıp amacı taşıyarak daha düşük boyutlu veriyi bir sonraki nöronun giriş katmanına taşır [36]. Bu yoğun (dense) katmanlarında gerçekleşen girdi verilerin sıkıştırılıp sonra tekrar düşük boyutlu iz düşümleri oluşturulduktan sonra orijinal veri ile karşılaştırılır. Verilen eşik değere göre yüksek bir hataya sahipse, anomali olarak ayrılır [17].

Tablo 10: Autoencoder Anomali Raporu

Anomali Sayısı	İşlem numarası ve tekrar adedi	Müşteri numarası ve tekrar adedi	Ürün numarası ve tekrar adedi	Genel İndirim Uygulanma Oranı	Genel Maximum Segment Oranı	Genel Maximum Satın Alma Frekansı Oranı
	8483 (2)	420 (17)	543 (14)			
146	7783 (2)	57 (13)	540 (12)	%100	%89,72	%92,46
					Yüksek Satış	Yüksek Frekans
	13799 (2)	375 (11)	192 (12)			

LSTM

RAM gereksinimi fazla olduğundan dolayı kullanılan sistem donanım yönünden yeterli olmamıştır. Sağlıklı sonuçlar alınamadığından bu algoritma değerlendirilme dışı bırakılmıştır.

SO-GAAL

SO-GAAL algaoritması için GAN (Generative Adversarial Network, Üretken Karşıt Ağlar) modeli oluşturulmuştur [37]. İlk aşamada üretici (generatorG) fonksiyonu için giriş ve ara katmanlarında ReLU kullanılırken çıkış katmanında Linear aktivasyon fonksiyonu kullanılmıştır. Üretici gerçek verileri kullanarak benzer ve sahte aykırı değerler üretir [38]. İkinci aşamada ayrıştırıcı (discriminator) fonksiyonu tanımlanır. Ayrıştırıcı, üretici tarafından oluşturulan sahte veriler üzerinde çalışır ve değerleri normal ve anormal olarak sınıflandırır [38]. Giriş ve ara katmanlarda ReLU aktivasyon fonksiyonu kullanılırken son katmanda Sigmoid kullanılmıştır.

Tablo 11: SO-GAAL Anomali Raporu

Anomali Sayısı	İşlem numarası ve tekrar adedi	Müşteri numarası ve tekrar adedi	Ürün numarası ve tekrar adedi	Genel İndirim Uygulanma Oranı	Genel Maximum Segment Oranı	Genel Maximum Satın Alma Frekansı Oranı
	3165 (8)	230 (307)	192 (67)			
1871	2444 (7)	424 (59)	1756 (37)	%99,73	%26,83	%86,42
					Ortalama Satış	Yüksek Frekans
	2648 (7)	80 (52)	527 (34)			

2.2.4 Diğer Özel Yöntemler

Özel birkaç anomali tespiti yöntemleri de incelenerek, verilerin özelliklerine ve analiz yapılacak senaryoya bağlı olarak etkinlikleri değerlendirilmiştir. Bu diğer yöntemler arasında Isolation Forest ve OCSVM gibi algoritmalar yer almaktadır. Aynı zamanda Isolation Forest algoritmasını Feature Bagging algoritması ile kullanılarak oluşabilecek değişimlerde gözlenmiştir.
Kullanılan Diğer Özel Anomali Tespit Algoritmaları:

Isolation Forest

Veri kümesi eğitim ve test olmak üzere ikiye ayrılır. Eğitim veri kümesi, yalıtılmış alt kümelere ayrılarak yapraklara kadar bu işlem devam edilir ve test veri kümesi üzerinde izolasyon yolu hesaplanır, böylece anomali skorları elde edilmiş olur [39]. Kısa yollar normal verileri temsil ederken uzun yollar ise aykırı değerleri temsil etmektedir. Diğer kullanılan algoritmalara göre %20 ile en yüksek anomali oranı bulunmuştur.

Tablo 12: Isolation Forest Anomali Raporu

Anomali Sayısı	İşlem numarası ve tekrar adedi	Müşteri numarası ve tekrar adedi	Ürün numarası ve tekrar adedi	Genel İndirim Uygulanma Oranı	Genel Maximum Segment Oranı	Genel Maximum Satın Alma Frekansı Oranı
	8716 (13)	230 (842)	192 (193)			
5821	7823 (9)	420 (131)	1756 (173)	%98,62	%35,28	%85,99
					Ortalama Satış	Yüksek Frekans
	7258 (8)	424 (125)	1830 (115)			

Feature Bagging

Temel algoritma olarak Isolation Forest kullanılmıştır. Bootstrap tekniği ile rastgele veriler seçilerek, veri kümesi alt kümelere ayrılmıştır. Alt küme sayısı veri kümesi boyutu ile aynı olacak şekilde ayarlanmıştır. Isolation forest algoritması her bir alt küme için çalıştırılıp anomali skorları elde edilmiştir. Tüm iterasyonlar için oluşturulan anomali skorlarının ortalaması alınıp aykırı değerlere ulaşılmıştır. Sadece Isolation Forest kullanılarak yapılan çalışma ile kıyaslandığında anomali sayısında %15,01 değerinde bir düşüş gözlemlenmektedir. Dolayısıyla aykırı değerlerin gerçek sayısı bilinmediği halde diğer algoritmalar ile kıyaslandığında Isolation Forest çok yüksek bir anomali oranı çıkarmıştır. Bu da performansı hakkında şüpheye düşürmektedir. Diğer özellikler açısından bakıldığında ise çok benzer sonuçlar alınması da Isolation Forest'ın gerçek aykırı değerleri doğru tespit etmiş olabileceğini ama bazı noktalarda benzerliklerden dolayı normal değerleri de seçip hataya düştüğü yorumunu yapmamıza olanak sağlar. Aynı zamanda buradan da Feature Bagging [40,

41] algoritmasının kullanımının, temel algoritmanın performansını ve güvenilirliğini iyileştirmede katkısı olduğu gözlenmektedir.

Tablo 13: Feature Bagging Anomali Raporu

Anomali Sayısı	İşlem numarası ve tekrar adedi	Müşteri numarası ve tekrar adedi	Ürün numarası ve tekrar adedi	Genel İndirim Uygulanma Oranı	Genel Maximum Segment Oranı	Genel Maximum Satın Alma Frekansı Oranı
	8613 (4)	230 (209)	192 (44)			
1455	7483 (4)	420 (35)	1830 (43)	%97,80	%34,50	%85,15
					Ortalama Satış	Yüksek Frekans
	8210 (3)	290 (29)	1756 (35)			

OC-SVM

OC-SVM algoritması alan bazlı bir yöntemdir, eğitilmiş model verileri örneklem uzayında konumlandırır ve orijine uzaklıklarını hesaplar. Orijine ne kadar yakınsa normal, karar sınırına yani hiper düzeleme ne kadar yakınsa anormal olarak sınıflandırılır [27]. Karmaşık veri yapılarında başarılı olmasına karşın sadece küresel çekirdek fonksiyonları ile çalışabilmesi sınırlı çalışma yetkinliği tanımaktadır [37].

Tablo 14: OC-SVM Anomali Raporu

Anomali Sayısı	İşlem numarası ve tekrar adedi	Müşteri numarası ve tekrar adedi	Ürün numarası ve tekrar adedi	Genel İndirim Uygulanma Oranı	Genel Maximum Segment Oranı	Genel Maximum Satın Alma Frekansı Oranı
	12030 (8)	230 (93)	540 (92)			
1455	13799 (5)	57 (81)	192 (91)	%93,33	%62,88	%80,82
					Yüksek Satış	Yüksek Frekans
	13798 (5)	420 (79)	543 (87)			

3. BULGULAR VE TARTIŞMA

Tablo 15: Maksimum Anomali Skoruna Sahip Verilerin Bilgileri

Algoritma Adı	Veri Kümesi Anomali Oranı	İşlem Numarası	Müşteri Numarası	Tarih	İndirim Kontrolü	Segment	Satın Alma Frekansı
AOM	-	-	-	-	-	-	-
Autoencoder	%0,50	11242	15	Şubat 2023	1	Yüksek	Yüksek
AvgkNN	%5	6155	230	Temmuz 2020	1	Yüksek	Yüksek
Feature Bagging	%4,99	5014	230	Mayıs	1	-	Yüksek
GMM	%5	1163	80	Ağustos	1	Ortalama	Yüksek
Isolation Forest	%20	261	420	Kasım 2020	0	Yüksek	Yüksek
kNN	%5	6155	230	Temmuz 2020	1	Yüksek	Yüksek
LOF	%8,96	261	420	Kasım 2020	0	Yüksek	Yüksek
LSCP	%0,05	376	481	Aralık	1	Ortalama	Yüksek
MCD	%2	2744	290	Kasım	1	-	Yüksek
OCSVM	%4,99	7074	272	2022	1	Yüksek	-
PCA	%1,79	427	58	Temmuz	1	Ortalama	Yüksek
SO-GAAL	%6,42	4225	66	Mart	1	-	Ortalama

En yüksek aykırı değer skoruna sahip veri noktasına göre tablo oluşturulmuştur.

Yöntemler bölümünde algoritmaların tek tek aykırı değerlerine bakılıp en çok tekrar eden 3 aykırı noktaya sahip veriler tablo haline getirilmiştir. Genel değerlendirmeler ve yorumlar ilgili bölümün altında yer almaktadır. Aykırı değerlerin geneline bakıldığında, anomalilerin büyük çoğunluğunda ortalama %95,58'nde indirim uygulandığını göstermektedir. Buradan indirim uygulandıysa anomalidir yorumunu yapmak hatalı bir yaklaşımdır. Aksi durumu Tablo 15 'de Isolation Forest ve LOF algoritmalarında en yüksek anomali skoruna sahip veride indirim uygulanmadığını göstermektedir. Fatura işlem numaralarına bakıldığında ise ilk üçte ortak olarak bulunan 7804, 13799, 11392 ve 12030 numaralı faturalar ortak özellik olarak dikkat çekmektedir. Çalışanların kontrol yapması durumunda bu 4 faturaya öncelikli olarak bakılması tavsiye edilebilir. Segment yüzdeliklerine bakıldığında ortalama ve yüksek segment üzerinde değişken oranlara sahiptir. Düşük segmente, şirkete düşük gelir getirisine sahip müşterilere uygulanan tüm algoritmalarda rastlanmamıştır. Ama bu anomali tespitinde sağlıklı bir belirteç olacağı anlamına gelmez, tek başına bu öncüle göre yorum yapmak doğru değildir. Her müşterinin satın aldığı ürün adedinin frekansının kategorik veriler üzerinde durumu, uygulanan tüm algoritmalarda genel ortalamasına bakıldığında %83,06 ile yüksek frekansta olanların aykırı değer olarak rastlanması ihtimali yüksektir. Etiketli veriler bulunmadığı için kesin bir kanıya varılamasa da satışta ve faturalandırılmada yaşanabilecek olası krizlerde hata sebebinin hızlı bulunmasına olanak sağlar. Yetkilinin olası anomaliye odaklanması ve geniş veri kümesini filtrelemesi kolaylaştırılır. Etiketli veriler bulunmadığından ve tarih verileri de çok değişkenlik gösterdiğinden dolayı bunun üzerine çıkarım yapmak doğru değildir. Perakende veri kümesi üzerinde aykırı değerlerin tespitinde yorumlanabilirlik başarısının en yüksek olduğu öncül ve anomali skorlarının tüm algoritmalarda toplandığı müşteri numaralarının incelenmesi öneri

olarak sunulabilir. İstatistiksel olarak diğer algoritmalarla birlikte aykırı değere sahip olması, yani şüpheli duruma alınabilecek müşteriler Tablo 16 da verilmiştir.

Tablo 16: Riskli Müşteriler

Müşteri Numarası	52	57	230	375	420	424
Anomali olma olasılığı	%0,32	%1,29	%18,78	%0,87	%2,83	%1,34

Kullanılan algoritmalarda tekrar sayısına ve anomali skorlarının yüksekliğine göre seçilmiş 6 adet müşteri yer almaktadır. Aynı zamanda aykırı değer tespit edilen 12 algoritma içinde 230 ve 420 müşteri numarasına sahip olanlar 8 algoritmada da yüksek tekrar sayılarına ulaşmıştır. Bu yüzden olası bir sorunda yetkililer için kontrol edilmesi önerilen iki müşteri olarak seçilebilirler.

4. SONUÇLAR

Anomalilerin birçok disiplinde ciddi sorun teşkil etmesiyle birlikte asıl kullanılan veri kümesi özelinde, KOBİ'lerin ihtiyaçları göz önüne alındığında iş akışının ve finansal durumlarının kontrolden çıkmaması adına dikkat edilmesi gereken bir problemdir. Bu amaç doğrultusunda satış anomalilerinin hızlı yakalanması hedeflenmektedir, KOBİ'lerin verimliliği ve iş akışının stabil ilerlemesi için Kolay.AI adlı yazılımda da bu konu üstünde durulmuştur [42]. Kaggle veri tabanından alınan perakende satış verileri üzerinden bazı anomali tespit algoritmaları kullanılarak performans değerlendirilmesi ve etiketsiz veriler için analiz sonuçlarının yorumlanması üzerine çalışılmıştır [43]. AOM, Autoencoder, AvgkNN, Feature Bagging, GMM, Isolation Forest, kNN, LOF, LSCP, MCD, OCSVM, PCA ve SO-GAAL teknikleri uygulanmış olup veri ön işleme, özellik çıkarımı ve algoritma sonuçlarının yorumlanması üzerine bilgi verilmiştir. Benzer veri kümelerinde anomali tespiti yaparken algoritma seçimi, yorumlama ve anlam çıkarımı aşamaları için fikir verme amacı taşımaktadır. Aynı zamanda veri kümesinin zenginleştirilmesi ve bu sayede oluşacak yeni değişkenlerin başarılı bir aykırı değer tespitinde önemli olduğu göz önüne getirilmiştir.

REFERANSLAR

[1] Zhao, Yue, Zain Nasrullah, and Zheng Li. "Pyod: A python toolbox for scalable outlier detection." *arXiv preprint arXiv:1901.01588* (2019).

[2] Hodge, Victoria, and Jim Austin. "A survey of outlier detection methodologies." *Artificial intelligence review* 22 (2004): 85-126.

[3] Fanaee-T, Hadi, et al. "Event and anomaly detection using tucker3 decomposition." *arXiv preprint arXiv:1406.3266* (2014).

[4] Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." *ACM computing surveys (CSUR)* 41.3 (2009): 1-58.

[5] Brax, Christoffer. *Anomaly detection in the surveillance domain*. Diss. Örebro universitet, 2011.

[6] ÖRNEK, Özlem, Eyyüp GÜLBANDILAR, and Ahmet YAZICI. "AKILLI FABRİKALARDAKİ OTONOM TAŞIYICILAR İÇİN BULANIK MANTIK TABANLI ANOMALİ TESPİTİ." *Eskişehir Osmangazi Üniversitesi Mühendislik ve Mimarlık Fakültesi Dergisi* 28.1 (2020): 53-61.

[7] Schölkopf, Bernhard, Alexander Smola, and Klaus-Robert Müller. "Nonlinear component analysis as a kernel eigenvalue problem." *Neural computation* 10.5 (1998): 1299-1319.

[8] Fujimaki, Ryohei, Takehisa Yairi, and Kazuo Machida. "An approach to spacecraft anomaly detection problem using kernel feature space." *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. 2005.

[9] Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Outlier detection: A survey." *ACM Computing Surveys* 14 (2007): 15.

[10] Yolaçan, Esra Nergis. *Learning from sequential data for anomaly detection*. Northeastern University, 2014.

[11] Goldstein, Markus, and Seiichi Uchida. "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data." *PloS one* 11.4 (2016): e0152173.

[12] Yılmaz, Erhan. *Makine öğrenmesi tabanlı kullanıcı davranış analizi ile bilgisayar sistemlerine giriş kayıtlarında anomali tespiti*. MS thesis. Ege Üniversitesi, Sosyal Bilimler Enstitüsü, 2022.

[13] Iivari, Albin. "Anomaly detection techniques for unsupervised machine learning." (2022).

[14] Kotsiantis, Sotiris B., Ioannis Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." *Emerging artificial*

intelligence applications in computer engineering 160.1 (2007): 3-24.

[15] Kriegel, Hans-Peter, Matthias Schubert, and Arthur Zimek. "Angle-based outlier detection in high-dimensional data." *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2008.

[16] Goldstein, Markus, and Andreas Dengel. "Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm." *KI-2012: poster and demo track* 1 (2012): 59-63.

[17] Nguyen, Nhan-Tam. *Unsupervised outlier detection in official statistics*. Vol. 57. Bank for International Settlements, 2022.

[18] Ding, Chris, and Xiaofeng He. "K-means clustering via principal component analysis." *Proceedings of the twenty-first international conference on Machine learning*. 2004.

[19] Brauckhoff, Daniela, Kave Salamatian, and Martin May. "Applying PCA for traffic anomaly detection: Problems and solutions." *IEEE INFOCOM 2009*. IEEE, 2009.

[20] Hubert, Mia, Peter J. Rousseeuw, and Stefan Van Aelst. "Multivariate outlier detection and robustness." *Handbook of Statistics* 24 (2005): 263-302.

[21] Hardin, Johanna, and David M. Rocke. "Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator." *Computational Statistics & Data Analysis* 44.4 (2004): 625-638.

[22] Hubert, Mia, and Michiel Debruyne. "Minimum covariance determinant." *Wiley interdisciplinary reviews: Computational statistics* 2.1 (2010): 36-43.

[23] He, Zengyou, Xiaofei Xu, and Shengchun Deng. "Squeezer: an efficient algorithm for clustering categorical data." *Journal of Computer Science and Technology* 17.5 (2002): 611-624.

[24] He, Zengyou, Xiaofei Xu, and Shengchun Deng. "Discovering cluster-based local outliers." *Pattern recognition letters* 24.9-10 (2003): 1641-1650.

[25] Gao, Zengan. "Application of cluster-based local outlier factor algorithm in anti-money laundering." *2009 International Conference on Management and Service Science*. IEEE, 2009.

[26] Yang, Xingwei, Longin Jan Latecki, and Dragoljub Pokrajac. "Outlier detection with globally optimal exemplar-based GMM." *Proceedings of the 2009 SIAM international conference on data mining*. Society for Industrial and Applied Mathematics, 2009.

[27] Grammatikis, Panagiotis Radoglou, et al. "An anomaly detection mechanism for IEC 60870-5-104." *2020 9th International Conference on Modern Circuits and Systems Technologies (MOCAST)*. IEEE, 2020.

[28] Janssens, Jeroen HM. "Outlier selection and one-class classification." (2013).

[29] Yang, Peng, and Biao Huang. "KNN based outlier detection algorithm in large dataset." *2008 International Workshop on Education Technology and Training & 2008 International Workshop on Geoscience and Remote Sensing*. Vol. 1. IEEE, 2008.

[30] Zhu, Rui, et al. "KNN-based approximate outlier detection algorithm over IoT streaming data." *IEEE Access* 8 (2020): 42749-42759.

[31] Zhao, Yue, et al. "LSCP: Locally selective combination in parallel outlier ensembles." *Proceedings of the 2019 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, 2019.

[32] Breunig, Markus M., et al. "LOF: identifying density-based local outliers." *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. 2000.

[33] Duan, Lian, et al. "Cluster-based outlier detection." *Annals of Operations Research* 168 (2009): 151-168.

[34] Papadimitriou, Spiros, et al. "Loci: Fast outlier detection using the local correlation integral." *Proceedings 19th international conference on data engineering (Cat. No. 03CH37405)*. IEEE, 2003.

[35] Chapelle, Olivier, Bernhard Scholkopf, and Alexander Zien. "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]." *IEEE Transactions on Neural Networks* 20.3 (2009): 542-542.

[36] Bank, Dor, Noam Koenigstein, and Raja Giryes. "Autoencoders." *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook* (2023): 353-374.

[37] Liu, Yezheng, et al. "Generative adversarial active learning for unsupervised outlier detection." *IEEE Transactions on Knowledge and Data Engineering* 32.8 (2019): 1517-1528.

[38] Creswell, Antonia, et al. "Generative adversarial networks: An overview." *IEEE signal processing magazine* 35.1 (2018): 53-65.

[39] Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation forest." *2008 eighth ieee international conference on data mining*. IEEE, 2008.

[40] Xu, Xiaodan, Huawen Liu, and Minghai Yao. "Recent progress of anomaly detection." *Complexity* 2019 (2019).

[41] Lazarevic, Aleksandar, and Vipin Kumar. "Feature bagging for outlier detection." *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. 2005.

[42] ŞEKER, Şadi Evren. "KOBİ'lere Özel Basit Yapay Zeka Çözümü: Kolay. AI." *YBS Ansiklopedi*. v.11, is.1, 2023.

[43] Sadi Evren SEKER. "Retail Data Set". *Kaggle*. 2023. https://doi.org/10.34740/KAGGLE/DS/3067824