**Online Retail Üzerinde RFM Analizi ile Müşteri Segmentasyonu**

RFM analizi, müşteri davranışlarını anlamak ve müşteri segmentasyonu yapmak için kullanılan bir yöntemdir. RFM, Recency (Yenilik), Frequency (Sıklık) ve Monetary Value (Mali Değer) kelimelerinin baş harflerinden oluşur ve müşterilerin bu üç temel özelliği üzerinden değerlendirilmesine dayanır.

RFM analizi genellikle bir veri setindeki her müşteri için bu üç özelliği içeren bir skorlama sistemini içerir. Her bir RFM bileşeni genellikle belirli aralıklara bölünerek, müşterilere puanlar atanır. Bu puanlar daha sonra bir araya getirilerek müşteriler belirli segmentlere ayrılır. RFM analizi, özellikle müşteriye özel pazarlama stratejileri geliştirmek ve müşteri sadakatini artırmak için kullanılır.

**Veri Özeti**

Veri seti, bir Birleşik Krallık merkezli çevrimiçi hediye mağazasında 01/12/2010 ile 09/12/2011 tarihleri arasında gerçekleşen işlemlerle ilgili verileri içerir.

**Identifiers**

**InvoiceNo:** Fatura numarası. Her işleme benzersiz olarak atanmış 6 basamaklı bir sayı. 'C' ile başlayanlar iptali gösterir.

**StockCode:** Ürün (madde) kodu. Her benzersiz ürüne benzersiz olarak atanmış 5 basamaklı bir sayı.

**CustomerID:** Müşteri numarası. Her müşteriye benzersiz olarak atanmış 5 basamaklı bir sayı.

**Kategorik Veri:**

**Country:** Ülke adı. Her müşterinin ikamet ettiği ülkenin adı.

**Description:** Ürün (madde) adı.

**Sayısal Veri:**

**InvoiceDate:** Fatura Tarihi ve Saati. Her işlemin oluşturulduğu gün ve saat.

**Quantity:** Her ürünün (madde) işlem başına miktarı.

**UnitPrice:** Birim fiyat. Sterlin cinsinden ürün fiyatı.

İlk olarak dataseti okuduk. Ardından analiz için belli bir tarih aralığı seçtik. Bu aralık 01/04/2011-09/12/2011 olarak seçildi. Sonrasında eksik değerleri kontrol ettik. CustomerID sütununda 90853 eksik değer olduğunu gördük. customerID identifier olduğundan analiz için önemli bir veridir analizin doğru olması açısından eksik değerleri drop ettik.

Veri ön işlemeye devam ederken metadata içerisinde iptal edilmiş işlemlerin bulunduğunu gözlemledik. Daha kesin bir analiz elde etmek için bu satırları kaldırdık. Devamındaysa yinelenen (duplicate eden) kayıtları tespit edip bunları da kaldırdık. Quantity ve Price sütunlarındaki negatif değerleri kontrol edip bu değerleri de filtreledik ve burada negatif değer bulunmadığını gördük.

Bu işlemler sonucunda datasetimiz temiz ve analize hazır hale getirdik.

**R (Recency) Hesaplama**

Recency bir müşterinin çevrimiçi mağazadan ne kadar süre önce satın alma yaptığının bir ölçüsüdür. Eğer bir müşteri Aralık 2011 ayında bir satın alma işlemi yaptıysa, Recency 0 olmalıdır. Eğer satın alma Kasım 2011'de yapılmışsa, Recency 1 olmalıdır ve bu şekilde devam etmelidir.

Bu hesaplama için dataframe’e R adında bir sütun ekledik. Bu sütun, indeks sütunundaki ay değerinin 12'den çıkarılmış değerini saklar (çünkü referans/analiz ayı Aralık'tır). Dolayısıyla, en güncel ay 0 değerini alır, ikinci en güncel ay 1 değerini alır, ve böyle devam eder.

**F (Frequency) Hesaplama**

Frequency hesaplamak için, F adında yeni bir sütun oluşturduk ve tüm satırlarını 0’dan başlattık. Ardından müşteri id’lerini tutan unique bir liste oluşturduk ve o müşteriye ait fatura sayısını F sütununa ekledik.

**M (Monetary Value) Hesaplama**

Monetary Value hesaplamak için, önce Revenue adında yeni bir sütun oluşturduk ve her bir item’ın unitprice ve quantitiy değerinin çarpımını saklamak için bu sütunu kullandık. Daha sonra M isimli bir sütun oluşturduk ve customerID’ler arasında dönerek her müşterinin ürettiği geliri bu sütuna ekledik.

Unique costumerID’lerini ve RFM değerlerini saklamak için yeni dataframe oluşturduk.

**K-Means Algoritması ile Müşteri Segmentleri Oluşturma**

K-means için optimal küme sayısını belirlemek için Elbow yöntemini kullandık. Bu yöntem farklı küme sayıları için modellerin performansını değerlendirir ve küme sayısının artmasıyla birlikte toplam karesel hatanın (within-cluster sum of squares, WCSS) nasıl değiştiğini analiz eder. Elbow yöntemi, WCSS'nin küme sayısıyla ilişkili grafiğini çizerek, küme sayısının artmasıyla birlikte WCSS'nin azalma eğilimini değerlendirir.

Elbow noktası, bu grafiğin dirsek benzeri bir bükülme noktasıdır. Bu nokta, küme sayısını artırmaya devam etmenin WCSS üzerindeki azalmada belirgin bir iyileşme sağlamadığı, yani artan küme sayısının ek fayda getirmediği noktadır. Bu, optimal küme sayısını belirlemede kullanılır, çünkü küme sayısı arttıkça modelin karmaşıklığı da artar, ancak aynı zamanda aşırı uyuma (overfitting) eğilimindedir. Elbow noktası, optimal dengeyi temsil eder.

WCSS grafiğini çizdirdiğimizde küme sayısının 3 olduğu noktada dirsek benzeri kırılma noktası oluştuğunu gözlemledik. Bu nedenle küme sayısını 3 olarak belirledik.

**Kümeleme Sonuçlarını Grafiklendirme**

Elbow metodu ile kümesi sayısını 3 olarak belirledikten sonra bununla RFM değerlerini kullanarak k-means modeli oluşturduk.

Sonrasında, dataframe'deki her gözlem için küme label’ını oluşturduk ve bunları bir sütunda sakladık.

Ardından kümeleri çizdirdik. Daha iyi görselleştirme için işaretleri birbirinden farklı ayarladık ve renklerin her label için değişmesini hue parametresiyle sağladık.

Grafikleme sonucunda 9 grafik oluşturulmuştur - Her biri Recency değerinin biri için-. X ekseni üzerinde Monetary Value, Y ekseni üzerinde ise Frequency bulunmaktadır. Her küme farklı bir renge ve farklı bir şekle sahiptir.

**Her Müşteri Segmentinin Sezgisel Olarak Açıklanması**

Bu noktada cluster center’ları kullanarak sezgisel yaklaşımda bulunduk.

0 için R değeri 4.38, F değeri 68.6 ve M değeri 1212'dir (Düşük Recency, Düşük Frekans ve Düşük Monetery Değerlere sahip müşteriler) - düşük değere sahip müşteriler.

1 için R değeri 4.00, F değeri 1296 ve M değeri 168158'dir (Düşük Recency, Yüksek Frekans ve Yüksek Monetery Değerlere sahip müşteriler) - sadık müşteriler, Sadakat Programına dahil edilmeli ve özel avantajlar sunulmalıdır.

2 için R değeri 5.25, F değeri 865 ve M değeri 39738'dir (Düşük Recency, Orta Frekans ve Orta Monetery Değerlere sahip müşteriler) - zaman zaman alışveriş yapan müşteriler - daha fazla promosyon ve teklif gönderilmelidir.

WCSS değerleri, noktaların ve küme merkezlerinin arasındaki kare mesafelerin toplamını hesaplayarak kümelerin kalitesini analiz etmek için kullanılır (WCSS ne kadar düşükse, kalite o kadar yüksektir).

**Agglomerative Algoritmasıyla Müşteri Segmentleme**

Elbow metodundan çıkardığımız sonuca göre 3 küme oluşturarak ve RFM değerlerini kullanarak Agglomerative Clustering modelini oluşturduk.

Burdan aldığımız sonuçları görselleştirmek için dendrogram çizdirdik. Görselleştirme sonucunda yukarı doğru gidildikçe birçok küçük kümenin bir araya gelerek daha büyük kümeleri oluşturduğunu gözlemledik.

**K-Means ve Agglomerative Clustering Sonuçlarını Karşılaştırma**

Bu iki kümelemenin sonucunu kıyaslamak için modellerin oluşturduğu labelları listeledik ve her bir kümenin boyutunu aldık.

Sonuca bakıldığında her iki algoritmanın üçer kümesi genel olarak aynı büyüklükte olduğunu gözlemledik. Bunların sonucunda algoritmaya bakılmaksızın elde edilen müşteri segmentlerinin genellikle aynı olduğunu gözlemledik.