T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

PERAKENDECİLİKTE VERİ MADENCİLİĞİ UYGULAMALARI VE SORUNLARI

GÜRKAN MOCAN

YÜKSEK LİSANS TEZİ
İSTATİSTİK ANABİLİM DALI

DANIŞMAN
PROF. DR. ALİ HAKAN BÜYÜKLÜ

İSTANBUL, 2016

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

PERAKENDECİLİKTE VERİ MADENCİLİĞİ UYGULAMALARI VE SORUNLARI

Gürkan MOCAN tarafından hazırlanan tez çalışması 11.05.2011 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Prof. Dr. Ali Hakan BÜYÜKLÜ

Yıldız Teknik Üniversitesi

Jüri Üyeleri

Prof. Dr. Ali Hakan BÜYÜKLÜ

Yıldız Teknik Üniversitesi _____

Yrd. Doç. Dr. Funda H. SEZGİN

Mimar Sinan Güzel Sanatlar Üniversitesi _____

Yrd. Doç. Dr. Doğan YILDIZ

Yıldız Teknik Üniversitesi _____

ÖNSÖZ

Gelişen bilgisayar teknolojileri özellikle perakende sektöründe işlev gören firmalara, müşteri verisi toplayabilme, kaydedebilme ve bu verileri işleyebilme gibi yetkinlikler kazandırmıştır. Bilgisayar teknolojilerindeki bu yetkinlik kuşkusuz pazarlama alanında fırsatlar sunmuştur. Günümüzde birçok perakende firması, müşterilerini bire-bir tanımak ve ihtiyaçlarına hizmet etmek istemektedir.

Kuşkusuz her bir müşterinin ayrı ayrı istek ve ihtiyaçlarına cevap vermek, özellikle çok büyük müşteri kitlesine sahip firmalar için pratikte mümkün olamayacaktır.

Bu nedenle perakendecilikte işlev gören bir firma Müşteri Ayrıştırma yöntemlerine ihtiyaç duymaktadır. Günümüzde müşterilerine onlara özel fırsatlar sunan ve onların ihtiyaçlarını bilen firmalar sektör içerisinde pazardan daha büyük pay aldıkları görülmektedir.

Bu tez içerisinde uygulamada hangi müşteri ayrıştırma yöntemlerine, ne zaman ihtiyaç duyulmaktadır? Uygulamada karşılaşılan sorunlar ve çözümler yeralamaktadır.

Emekleri ve desteklerinden dolayı Prof. Dr. Ali Hakan Büyüklü ve teze örnek veri (örnek ve sadece kodlanmış verilerden oluşmaktadır) sunan Boyner Büyük Mağazacılık A.Ş' ye, bilime olan katkılarından dolayı sonsuz teşekkürlerimi sunarım.

Ocak, 2011

Gürkan MOCAN

İÇİNDEKİLER

Sayfa

KISALTMA LİSTESİ viii

ŞEKİL LİSTESİ ix

ÇİZELGE LİSTESİ x

ÖZET xi

ABSTRACT xiii

BÖLÜM 1

GİRİŞ 1

1.1 Literatür Özeti 1
1.2 Tezin Amacı 1
1.3 Hipotez 2

BÖLÜM 2

BİLGİ KEŞİF SÜRECİ (BKS) 3

2.1 Veri Madenciliği Metodolojileri 3
2.1.1 Akademik Araştırma Modelleri 4
2.2 Endüstriyel Modeller 5
2.2.1 CRISP-DM 5
2.2.1.1 İşi Anlama 5
2.2.1.2 Veriyi Anlama 6
2.2.1.3 Veriyi Hazırlama 6
2.2.1.4 Modelleme 7
2.2.1.5 Değerlendirme (evaluation) 7
2.2.1.6 Yayılım 8
2.3 Melez Modeller 8
2.3.1 SEMMA (Sample; Explore; Modify; Model; Assess) 10
2.3.1.1 SEMMA Sürecinin Adımları 10

BÖLÜM 3

KÜMELEME 13

3.1 Kümeleme Tanımı 14
3.2 Kümeleme Analizinde Veri Tipleri 16
3.2.1 Uzaklık (mesafe) Ölçüleri 16
3.2.1.1 Veri Matrisi 16
3.2.1.2 Başkalık Matrisi 17
3.2.1.3 Aralıklı Ölçek Değişkenler 17
3.2.1.4 Mesafe Ölçüm Yöntemleri ve Standardizasyon 17
3.2.2 İkili Değişkenler için Mesafe Ölçütleri 19
3.2.3 Kategorik Değişkenler için Mesafe Ölçütleri 21
3.2.4 Sıralı Değişkenler için Mesafe Ölçütleri 22
3.2.5 Oran Ölçekli Değişkenler için Mesafe Ölçütleri 23
3.2.6 Karışık Tipli Değişkenler için Mesafe Ölçütleri 24
3.2.7 Vektör Gözlemler 25
3.2.7.1 Pearson Korelasyon Ölçüsü 26
3.2.7.2 Genişletilmiş Jaccard Ölçüsü (Tanimoto Katsayısı) 26
3.2.7.3 Dice Katsayı Ölçüsü 26
3.2.7.4 Değerlendirme Kriter Ölçüleri 26
3.2.7.5 İçsel Nitelik Kriteri 27
3.2.7.6 Hata Kareler Toplamı (SSE) 27
3.2.7.7 Diğer Minimum Varyans Kriteri 27
3.2.7.8 Serpilme Kriteri 28
3.2.7.9 İz Matris Kriteri 28
3.2.7.10 Determinant Kriteri 28
3.2.7.11 Değişmezlik Kriteri 29
3.2.7.12 Condorcet Kriteri 29
3.2.7.13 C-Kriteri 29
3.2.7.14 Kategori Fayda Metriği 29
3.2.7.15 Sınır Kesim Metriği 29
3.2.7.16 Dışsal Nitelik Kriterleri 30
3.2.7.17 Ortak Bilgi Temelli Ölçüm 30
3.2.7.18 Kesinlik-Geri Çağırma Ölçüsü 30
3.2.7.19 Rand İndeks 30

BÖLÜM 4

ALGORİTMA ve TEKNİKLER 31

4.1 Kümeleme Algoritmaları 31
4.2 Hedef Fonksiyon (Bölümlendirme Kriteri) Temelli Kümeleme 32
4.2.1 K-Ortalamalar 32
4.2.1.1 Algoritma 34
4.2.2 K-Medoid Algoritması 36
4.2.3 PAM (medoid etrafında bölümlendirme) 37
4.2.4 Büyük Veritabanları Bölümlendirmeli Kümeleme: CLARANS 38
4.3 Hiyerarşik Yöntemler 38
4.3.1 Birleştirici Hiyerarşik Kümeleme 39
4.3.2 Ayırıcı Hiyerarşik Kümeleme 39
4.4 Yoğunluk Temelli Kümeleme 42
4.5 Izgara Temelli Kümeleme (GRID) 43

4.6 Model Temelli Kümeleme ... 43

BÖLÜM 5

MÜŞTERİ İLİŞKİLERİ YÖNETİMİ ve PAZARLAMADA VERİ MADENCİLİĞİ YÖNTEMLERİ ... 44

5.1 Yaşam Döngüleri ... 44
5.1.1 Müşteri Yaşam Döngüsü ... 45
5.1.2 Müşteri Yaşam Döngüsü İşletme Süreçleri ... 49
5.1.3 Müşteri Kazanım ... 49
5.1.4 Aktivasyon ... 50
5.1.5 İlişki Yönetimi ... 50
5.1.6 Elde Tutma ... 50
5.1.7 Geri Kazanım ... 50
5.1.8 Müşterinin Yaşam Döngüsü: Yaşam Evreleri ... 52

BÖLÜM 6

PAZARLAMADA SAĞKALIM ANALİZİ ... 55

6.1 Müşteriyi Elde Tutma ... 56
6.2 Hazard ... 57
6.2.1 Risk(Hazard) Fonksiyonu Örnekleri ... 58
6.2.1.1 Sabit Risk Fonksiyonu ... 58
6.2.1.2 Banyo Küveti Şekilli Risk Fonksiyonu ... 59
6.2.1.3 Gerçek Yaşamdan Örnek ... 59
6.3 Sansürleme ... 60

BÖLÜM 7

PERAKENDE' DE VERİ MADENCİLİĞİNDE KULLANILACAK VERİ YAPISI ... 73

7.1 Ürün Verisi ... 73
7.2 İşlemsel Verileri ... 74
7.3 Müşteri Demografik Verileri ... 75
7.4 Müşteri İlişkileri Verileri ... 75
7.5 Ek Veriler ... 76
7.6 İşlem Bazında Veri Modeli ... 76
7.7 Uygulama Verisi ... 77
7.8 Müşteri Bazında Veri Modeli ... 79

BÖLÜM 8

MÜŞTERİ SEGMENTASYONU ... 82

8.1 Segmentasyonun Sonuçları ... 82
8.2 Başarılı Bir Segmentasyonun Özellikleri ... 83

BÖLÜM 9

SEGMENTASYON MODELLERİ ... 84

9.1 Değer Temelli Segmentasyon ... 84
9.2 Davranışsal Segmentasyon ... 85

9.3 Eğilim Bazlı Segmentasyon 85
9.4 Sadakat Segmentasyonu 86
9.5 Sosyo-Demografik ve Yaşam-Evresi Bazlı Segmentasyon Modeli 88
9.6 İhtiyaç Bazlı Segmentasyon 89

BÖLÜM 10

SGP (Sonzamandardalık, Geliş Sıklığı, Parasal) – RFM ANALİZİ 91

10.1 SGP Bileşenlerinin Sürekli SGP Skoruna Dönüştürülmesi 93
10.2 Gerçek Müşteri Verileri SGP Modeli 94
10.3 SGP Segmentasyon Prosedürü 95
10.3.1 Veri Elde Edinim 95
10.3.2 Segmente Edilecek Kitlenin Seçimi 95
10.3.3 Veri Hazırlığı ve S, G ve P Ölçümlerinin Hesaplanması 95
10.3.4 SGP Hücrelerinin Oluşumu, Kantil Gruplanma 95
10.3.5 Kümeleme ile SGP Segmentlerinin Keşfedilmesi 96
10.4 Gerçek Veriler- Örnek Olay İncelemesi 97
10.4.1 K-Ortalamalar Kullanımı 97
10.4.1.1 Profiller 100
10.4.2 Yaşam Evrelerine Göre SGP Model Segmentasyon 101
10.4.2.1 Profiller 102

BÖLÜM 11

SONUÇ VE ÖNERİLER 104

ÖZGEÇMİŞ 109

KISALTMA LİSTESİ

BKS(KDD)	Bilgi Keşif Süreci (BKS)
CLARANS	Clustering LARge Applications
CRISP-DM	Cross Industry Standard Process for DM
CRM (MİY)	Müşteri İlişkileri Yönetimi
EM	Expectation Maximization
OYF	Olasılık Yoğunluk Fonksiyonu
PAM	Partition Around Medoid
SEMMA	Araştırma, Tanımlama, Modelleme, Değerlendirme
SGP (RFM)	Sonzamandardalık-Sıklık-Parasal (Recency-Frequency -Monetary)
SKU	Stock-keeping units
UPC	Uluslararası ürün kodu
VM (DM)	Veri Madenciliği (Data Mining)

ŞEKİL LİSTESİ

Sayfa

Şekil 2.1 Bilgi keşif süreç tasarımı 3
Şekil 2.2 CRISP-DM Süreç Evreleri 5
Şekil 2.3 Melez modellerin yapısı 9
Şekil 2.4 SEMMA süreç tasarımı 12
Şekil 3.1 İkili değişkenler mesafe ölçüsü hesaplama örneği 21
Şekil 4.1 K-Ortalamalar Bölümlendirme 35
Şekil 4.2 Temsili gözlem ile yeni gözlemin yer değiştirmesi 37
Şekil 4.3 Dendogram, verideki görsel yapı 39
Şekil 4.4 A ve B kümelerine ilişkin mesafe ölçüm yöntemleri 41
Şekil 5.1 Müşteri yaşam döngüleri ve müşteri yaşam evreleri 45
Şekil 5.2 Farklı evreler içerisindeki müşteri yaşam döngü süreci 48
Şekil 5.3 Müşteri yaşam döngüsü içerisinde yer alan işletme süreçleri 51
Şekil 5.4 İşletme aksiyonları 51
Şekil 6.1 Türkiye Hayat ve Hayat Annüite Tablolarının Oluşturulması Projesi Sonuç Raporuna, 1927-2000 nüfus sayımlarına göre Türkiye Hayat Tabloları .. 58
Şekil 6.2 Gerçek değerlerle risk fonksiyonu 60
Şekil 6.3 Farklı zamanlarda başlayan müşteri grupları, bazı müşteriler hala aktif oldukları için sansürlenmiştir. 61
Şekil 7.1 Ürün verisi kırılımları 74
Şekil 9.1 Değer bazlı segmentasyon ile churn skoru düşük müşterilerin dağılımı, riskde değer segmentasyonu 86
Şekil 9.2 Sadakat Segmentleri 87
Şekil 9.3 Müşteri yaşam evreleri banka örneği 89
Şekil 10.1 hücre atamaları 92
Şekil 10.2 Kantillerle gruplanan SGP hücreleri 93
Şekil 10.3 Gerçek veriler ile SGP müşteri skor kıyaslaması 95
Şekil 10.4 Segmentlerin uzamsal gösterimi 100
Şekil 10.5 Subjektif prototipler ile seçilen segmentlerin ciro ve müşteri sayısı 100
Şekil 10.6 Segment yaş grup dağılımları 101
Şekil 10.7 Segment cinsiyet kırılımları 101
Şekil 10.8 “18-25” yaş grubu 102
Şekil 10.9 “26-35” yaş grubu 102
Şekil 10.10 “36-45” yaş grubu 103
Şekil 10.11 “46 üzeri” yaş grubu 103

ÇİZELGE LİSTESİ

Sayfa

Çizelge 3.1 İkili değişkenler için çapraz tablo 20
Çizelge 3.2 Değişken tiplerine göre 4 gözleme ait örnek veriler 21
Çizelge 5.1 Yaşam döngüsü evreleri ve davranış alışkanlıkları 53
Çizelge 6.1 Müşteri olma süreleri 62
Çizelge 6.2 Zaman periyodunda müşteri takibi 62
Çizelge 6.3 Risk olasılıkları 63
Çizelge 6.4 Müşteri Aylık harcama tutarları/kategorik 65
Çizelge 6.5 Aylık Harcama ve İşlem bilgileri/kategorik 66
Çizelge 6.6 Pasiflik Sınırları 70
Çizelge 6.7 Pasif Müşteri Bilgileri / Segment 71
Çizelge 6.8 Aktif Müşteri Bilgileri / Segment 71
Çizelge 7.1 Örnek Müşteri Transaction Tablosu (SAS library) 78
Çizelge 7.2 Müşteri Demografik Verileri 79
Çizelge 7.3 Gerçek bir perakende firmasına ait verilerden oluşturulan datamart yapısı 80
Çizelge 10.1 “proc fastclus” prosedürü ile rastgele başlangıç prototipler ile gerçekleşen kümeleme 96
Çizelge 10.2 Subjektif bilgiler ile başlangıç prototiplerinin oluşturulması sonucu elde edilen küme değerleri 99

ÖZET

PERAKENDE SEKTÖRÜNDE VERİ MADENCİLİĞİ UYGULAMALARI ve SORUNLARI

Gürkan MOCAN

İstatistik Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Prof. Dr. Ali Hakan BÜYÜKLÜ

Veri madenciliği işletmeler için sahip oldukları veriyi tanıma (tezin içeriği nedeniyle, bizim için müşteriyi tanıma) yeni ve değerli bilgileri keşfedebilme açısından çok önemlidir. Müşteri davranışlarının ve ihtiyaçlarının ortaya çıkarılması, aynı zamanda müşterilere yeni ihtiyaçların ortaya çıkarılmasını da sağlamaktadır. Perakende firmaları için veri madenciliği, geniş veri setleri içerisinden anlamlı ve ilginç yapıların ortaya çıkarılmasına yönelik yapılan araştırma ve analizlerdir.

İşletmeler, teknoloji sayesinde kaydı giderek kolaylaşan ve bu bağlamda giderek artan ve genişleyen veri kümelerine sahip olmuşlardır. Diğer taraftan da bu artan karmaşıklıkta gerçek bilgiye ulaşmak daha da zorlaştırmaktadır. Günümüzde veri madenciliği, bu ve benzeri zorluklar yaşayan tüm sektörlerin işini kolaylaştırmıştır. Akademik alanda ve sektörde bu anlamda birçok algoritma ve yöntem geliştirilmiştir. Perakendecilik sektöründe hizmet veren firmaların yoğun olarak kullandıkları bu algoritmaların uygulamadaki yeterliliklerini ve aksaklıklarını da ortaya çıkarmak, daha anlamlı bilgilerin ortaya çıkmasını sağlayabilmektedir. Geniş veri yapıları içerisinden anlamlı ve ilginç bilgilerin keşfedilmesinde teorik olarak birçok yaklaşım ve algoritmanın bulunmaktadır. Bu yöntem ve algoritmaların sonuçlarının ve işleyişlerinin uzman görüşleri olmadan çalıştırılması ve işletme gereksinimlerinin dikkate alınmaması gibi durumlarda; aslında çok anlamlı olmayan ve uygulamada bir takım engellerle karşılaşılan yararsız bilgiler

ortaya çıkabilmektedir. Sonuç olarak, subjektif faktörlerin bu algoritmalara ne şekilde etki ettiği ve bu etkilerin modele nasıl yansıtılabileceğini ifade eden öneriler getirilmeye çalışılacaktır.

Anahtar Kelimeler: Veri Madenciliği, RFM, modelleme, kümeleme, müşteri segmentasyon, churn, perakendecilikte veri madenciliği, kümeleme algoritmaları

ABSTRACT

DATA MINING TECHNIQUES AND ITS PROBLEMS IN RETAIL COMPANIES

Gurkan MOCAN

Department of STATISTICS

MSc. Thesis

Advisor: Prof. Dr. Ali Hakan BÜYÜKLÜ

Data mining is most useful in an exploratory analysis scenario in which there are no any predetermined information about what will constitute an "interesting" outcome. Data mining is the search for new, valuable, and nontrivial information in large volumes of data. It is also a cooperative effort of humans and computers. Best results are achieved by considering the knowledge of human experts in describing problems and goals with the search capabilities of computers.

For the retail companies, data mining is the exploration and analysis of large quantities of data in order to discover and describe meaningful patterns and rules. Data mining gives a better understanding of retailers' customers about their needs and behaviors.

Capabilities of both generating and collecting data have been increasing rapidly. Many enterprises acquire legacy databases as a result of the long history of information technology development. This is also help retailers to store and collect data history about their customers.

Data mining provides us to find meaningful and interesting patterns hidden in a large data sets. There are many methods and algorithms to find meaningful patterns in large data sets. It is also important explaining assumptions and lack of abilities of the methods and algorithms.

It is needed to balance computer capabilities with the human experience. And also it is important to consider retailers' needs.

Key words: Data mining, RFM, modeling, clustering, customer segmentation, churn, data mining for retail companies, clustering algorithms

BÖLÜM 1

GİRİŞ

1.1 Literatür Özeti

Veri madenciliği bilgisayar ortamında kayıtlı geniş veri setlerinin analiz edilmesidir. Örneğin bir market veya mağazada, müşterilerin alışverişleri ile büyük miktarda veri setleri meydana gelmektedir. Barkod sistemleri ile perakende firmaları alışveriş işlemlerini daha hızlı sağlayabilmekte ve bilgisayar kullanarak ürün fiyatlarını daha doğru tanımlayabilmektedirler. Böylece hem finansal açıdan karlılıklarını ve maliyetlerini daha doğru kontrol edebilmektedirler. Toplanan bu bilgiler aynı zamanda veri madenciliği için de bilgi kaynağı oluşturmaktadır.

Veri madenciliği bankalar tarafından kredi kartı müşterilerinin değerlendirilmesinde, sigorta ve telekomünikasyon firmaları için kötü kullanımın keşfedilmesinde, gsm firmaları için muhtemel kayıp müşterilerin belirlenmesinde, üretici firmalar için kalite kontrol süreçlerinde kullanılmaktadır.

1.2 Tezin Amacı

Veri madenciliği genellikle geniş veri kümelerinde istatistiksel analizleri ve/veya yapay zeka analizlerini içermektedir. İstatistiksel analizler, bir hipotez sunarak onun test edilmesi yaklaşımını izlerken, veri madenciliği yeni bilgi keşfi ile ilgilenmektedir.

Veri madenciliği perakende sektöründe bir çok problemde kullanılmaktadır;

- Müşteri Profilleme: Değerli müşteri kümelerinin belirlenmesi
- Hedef Kitle: Değerli müşterilerin karakteristik özelliklerinin belirlenmesi
- Müşteri Sepet Analizi: Müşteri ürün alışverişlerinin belirlenmesi

Perakende sektörü günümüzde rekabetin oldukça fazla olduğu bir sektördür. Bu rekabet ortamında yeni müşteri kazanmak ve bu müşterileri elde tutmak oldukça zor bir hale gelmiştir. Dolayısı ile, perakende firmaları müşterilerini tanıyıp, farklılaştırarak, ürün ve hizmetlerini ona göre özelleştirmek istemektedirler.

Perakende firmalarının farklı sosyo-ekonomik statüde müşterileri olsa dahi klasik pazarlama aksiyonlarına göre bire-bir pazarlama aksiyonlarına ağırlık vermektedir. Firma sadece tek bir sosyo-ekonomik statüde olan müşteri kitlesine hitap etse bile, rekabet içerisinde müşterilerini farklılaştırması ve onlara özel hizmet sunması gerekmektedir. Müşteri kitlesi aynı sosyo-ekonomik statüde olsa dahi, firmanın farklı ürün veya hizmet gamına hitap ediyor olabilir.

Perakende sektöründe özellikle çok farklı ürün gamı olan, çok farklı markaya sahip olan ve çok farklı müşterilere hitap eden firmalar için veri madenciliği çok önemli bir araçtır.

Bölüm 1' de veri madenciliği metodolojileri ve bilgi keşif süreçleri değerlendirilmiştir. Bu süreçler, madenciliğe başlamadan önce ve başladıktan sonra yapılması gereken süreçlerdir.

1.3 Hipotez

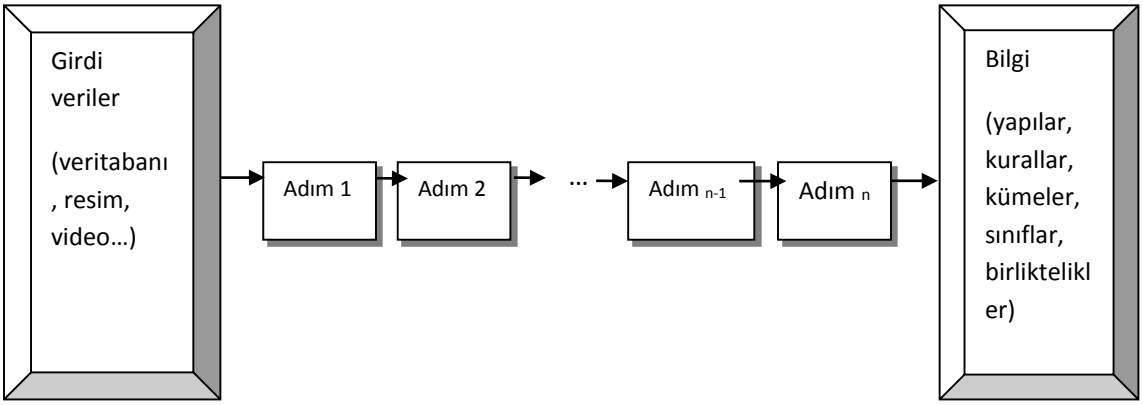
Gelişme bölümünde, müşteri ayrıştıma teknikleri ve algoritmaları, tekniklerin artı ve eksi yönleri, ayrıştırmada kullanılan mesafe ölçüm metrikleri, kümeleme yöntemleri açıklanmaktadır. Perakende ve diğer sektörlerde müşteri ayrıştırma yöntemleri, müşteri yaşam döngüleri, churn süreçleri ve SGP modelleri incelenmiş olup, uygulamaları gerçekleştirilmiştir. Ayrıca perakende veri yapıları ve bu veriler ile müşteri ayrıştırma yöntemi için belirlenen değişken yapıları yeralmıştır.

Sonuç olarak, perakende firmalarında müşteri ayrıştırma yönteminin sübjektif değerlendirmeler ile çok başarılı ve anlamlı bilgilere ulaşılabileceği; bu keşif sürecinde bir çok yöntemin birlikte değerlendirilmesinin önemi ortaya çıkarılmıştır. Aynı zamanda süreç içerisine firma gereksinimleri ve kısıtları da gözetilmesi gerekmektedir.

BÖLÜM 2

BİLGİ KEŞİF SÜRECİ (BKS)

Amaç ne kadar da veri içerisindeki faydalı bilgiye ulaşmak olsa da, genel yaklaşımı anlamak oldukça önemlidir. Bu nedenle süreci belirlemek yeni bilgiye erişmede oldukça kolaylık sağlar. Ortak bir çatıda bilgi keşif sürecini anlamak adına bazı süreç modelleri çıkarılmıştır. Bu modeller herhangi bir iş projesini yürütme ve planlamasa BKS' yi daha iyi anlamayı ve yol haritasını oluşturmayı sağlar. Bu projenin veya çalışmanın sonucunu kabulünü ve projeyi daha iyi algılamayı sağlar. Ayrıca maliye ve zaman faydası sağlar.



Şekil 2.1 Bilgi keşif süreç tasarımı

2.1 Veri Madenciliği Metodolojileri

Veri madenciliğini sistematik olarak yönetebilmek için genel süreçler takip edilmektedir. Bunlardan biri veri madenciliği çalışmalarında bir dizi adımları içeren Endüstri Standart Sürecidir (CRISP), diğeri de SAS' a özgü SEMMA' dır.

2.1.1 Akademik Araştırma Modelleri

90' ların ortalarına doğru şekillenmeye başlayan VM alanları ile, araştırmacıların karmaşık bilgi keşif dünyasında VM araçlarının kullanıcılarına rehber olan bir takım çoklu adımlı prosedürler geliştirmeye başladılar. Bunlardan biri Fayyad vd. dokuz adım modeli:

Uygulama alanını anlama ve geliştirme: Bu adım keşfedilmiş bilginin son kullanıcısının hedefini ve ilgili öncül bilginin öğrenilmesidir.

Hedef veri kümesinin oluşturulması: Veri madencisi keşif görevinde kullanacağı verileri ve değişkenler kümesini seçer. Bu adım birtakım sorgularla mevcut veri kümesi içerisinden istenilen örnek kümeleri sorgulanması sürecidir.

Veri temizleme ve hazırlık: Sonuçları etkileyecek sapan değerlerin tespiti, gerekiyorsa silinmesi. Eksik değerlerin ve hataların tespiti.

Boyut indirgeme ve izdüşümü: Anlamlı değişkenlerin tespiti ve veri ve değişkenlere ilişkin birtakım dönüşümlerin gerçekleştirilmesi.

VM görevinin seçimi: İlk adımdaki amaca uygun özellikli VM yönteminin seçilmesi, sınıflama, regresyon, kümeleme gibi.

VM Algoritmasının seçimi: Veri madencisi veri içerisindeki yapıları araştırmak için yöntemleri seçer ve kullanılacak yöntemlerin hangi model ve parametrelerinin uygun olduğunun kararını verir.

Veri Madenciliği: Trendler, sınıflama kuralları, karar ağaçları, regresyon modelleri gibi veri içerisindeki yapılar ortaya çıkar.

Model ve yapıların yorumu: Ortaya çıkan model ve yapıların görüntülenmesi.

Keşfedilen bilginin birleştirilmesi: Keşfedilen bilginin sisteme entegrasyonu, ilgili kişilere raporlanması ve belgelenmesi. Bu aşama ayrıca daha önceki bilinen bilginin doğruluğunu ve potansiyel çelişkilerinin çözümünü de içermektedir.

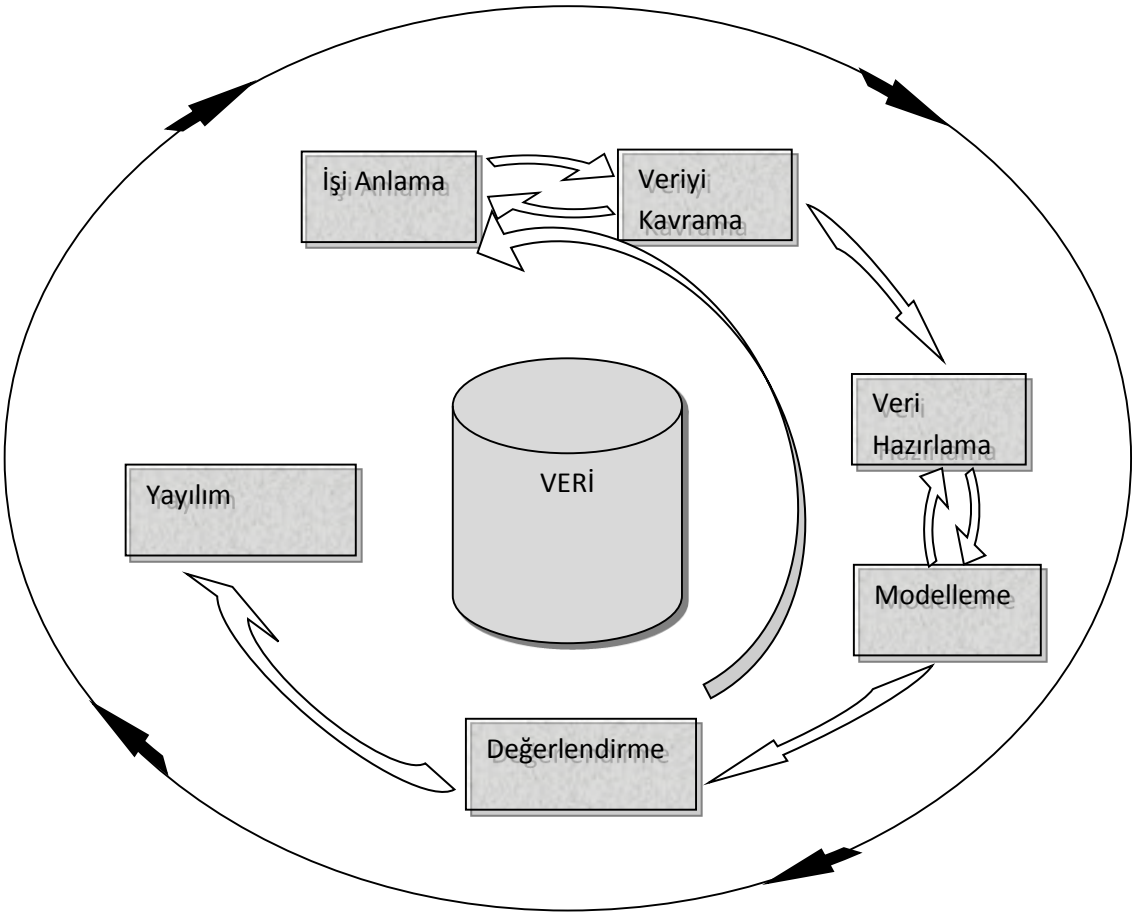
Bu model teknik anlamda veri analizi için detaylı bir açıklama sunabilmektedir fakat, iş dünyasına uygulama tarafında eksiklikleri mevcuttur. Yinede bu model kendinden sonraki modeller için temel unsurları barındırmıştır.

Kullanım alanları: Mühendislik, tıp, üretim, yazılım geliştirme ve elektronik ticaret.

2.2 Endüstriyel Modeller

2.2.1 CRISP-DM

Oldukça yaygın kullanılan veri madenciliği çalışmasının hayat döngüsü 6 ana evrede gerçekleşir. Şekil 2.2 veri madenciliği sürecinin evrelerini göstermektedir.



Şekil 2.2 CRISP-DM Süreç Evreleri

2.2.1.1 İşi Anlama

Veri madenciliği çalışmasında, çalışmanın amaçları ve gereksinimlerini işletme perspektifinde değerlendirilir. Birkaç adımda detaylandırılmaktadır,

- **İş amaçlarının tanımı:** Mevcut veya gelişen problemlerin iş dünyasının bakış açısında değerlendirilmesi
- **Durum değerlendirmesi:** Projenin gereksinimleri, varsayım ve kısıtları, mevcut riskleri ve kar-zarar analizleri gibi proje detayları değerlendirilir.

- **Veri madenciliğinin amaçlarının tanımı:** Bir iş hedefi iş terminolojisindeki amaçları ifade eder. Veri madenciliği de teknik olarak proje hedeflerini ifade etmektedir.
- **Proje planının oluşumu:** Veri madenciliği ve aynı zamanda iş hedeflerinin başarısı için amaçlanan planın tanımı

2.2.1.2 Veriyi Anlama

Temelde veri toplamayla başlar. Veri hakkındaki ilk izlenimler ve veri kalitesindeki sorunlar bu adımda değerlendirilir. Özetle şu adımları takip eder,

- İlk verinin toplanması: Proje kaynaklarında listelenen verinin elde edilmesi
- Veri tanımı: Veri yapısına dair genel bakış
- Verinin araştırılması: Bir takım sorgulamalar, grafiksel görünüm ve raporlama gibi veri hazırlığı öncesinde verinin araştırılmasıdır. Çıkan sonuçlar direk olarak veri madenciliği hedefine çıkarımlar sunabilir
- Veri kalitesinin doğrulanması: Veri gerçekten tam mı? Eksik değerler bulunmakta mı? Eğer varsa ne kadar sıklıkla?

2.2.1.3 Veriyi Hazırlama

Bu adımda veri setinin modelleme öncesinde hazır hale gelmesine yönelik yapılan veri temizleme, veri transformasyonları, manipülasyonları gerçekleşmektedir. Şu aşamalarda değerlendirilebilir,

Veri seçimi: Analiz için kullanılacak verilerin seçilmesi. Bu kriter, veri madenciliği amacı, teknik kısıtlar ve kalite ile ilintilidir.

Veri temizliği: Belirlenmiş analiz yöntemleri ile veri kalitesini istenilen seviyeye ulaştırmak amaçlanmıştır.

Veri mimarisinin inşası: Türetilmiş değişkenlerin, yeni kayıtların oluşturulması ve mevcut değişkenlerdeki değerlerin manipülasyonu gibi veri hazırlığında yapısal süreçleri içermektedir.

Veri entegrasyonları: Çoklu tabloların birleştirilmesi, yeni kayıt ve değerlerin oluşturulması

Verinin formatlanması: Verinin modellemeye girebilmesi için belirli gereksinimlere ihtiyacı olabilir. Bu modelleme aracı veya modelleme algoritmalarının gereksinimleridir.

2.2.1.4 Modelleme

Bu aşamada bir takım modelleme tekniklerinin seçimi ve uygulaması gerçekleşmektedir. Bazı metotlar belirli formatta girdi verisi gereksinimi duyabilir, bu nedenle bir önceki adıma tekrar bir iterasyon gereksinimi duyulabilmektedir. Bu adım,

Modelleme tekniklerinin seçimi: Uygulanacak gerçek modelleme tekniğinin seçilmesi ve varsayımlarının doğrulanması.

Test tasarımın oluşumu: Model inşası öncesinde, modelin kalitesi ve geçerliliğinin testi için bir mekanizma veya prosedür oluşturulması gerekmektedir.

Modelin oluşumu: Bir veya daha fazla model oluşturmak için hazırlanan veri kümesinin modelleme aracı ile çalıştırılması

Oluşturulan modelin değeri: Elde edilen veri madenciliği sonuçlarının teknik olarak değerlendirilmesi süreçlerinden oluşur.

2.2.1.5 Değerlendirme (evaluation)

Veri analizi bakış açısından yüksek kaliteye sahip model veya modeller oluşturulduktan sonra, İş dünyası bakış açısından model değerlendirilir. Model inşası için çalıştırılan adımların gözden geçirilir. Ana amaç yeterince gözlemlenmeyen önemli iş meselelerinin olup olmadığının saptanmasıdır. Anahtar adımlar,

- **Sonuçların değerlendirilmesi:** Bu adımda, modelin genelliği ve doğruluğuna dair faktörlerle ilgilenilir. İşin amacını karşılayan modelin değerlendirilmesi ve modelin neden eksik taraflarının olduğu bazı iş sebeplerinin olup olmadığının araştırılması. Diğer taraftan bütçe ve zaman dahilinde uygulamanın gerçek hayatta test edilmesi.
- **İşlemin gözden geçirilmesi:** Sonuç olarak ortaya çıkan model tatmin edici olmalı ve ticari ihtiyaçları karşılamalı. Veri madenciliği bağlamında daha derin kontrol sağlandığı, gözden kaçan faktör veya görevlerin olup olmadığı araştırılır. Bu aynı zamanda modelin doğru inşa edildiğinin sigortasıdır.
- **Sonraki adımların belirlenmesi:** Proje bu noktada ya bitirilecek ve “yayılım” adımına geçilecek yada ilave iterasyonlara veya yeni bir veri madenciliği projesinin başlamasına öncülük edecek.

2.2.1.6 Yayılım

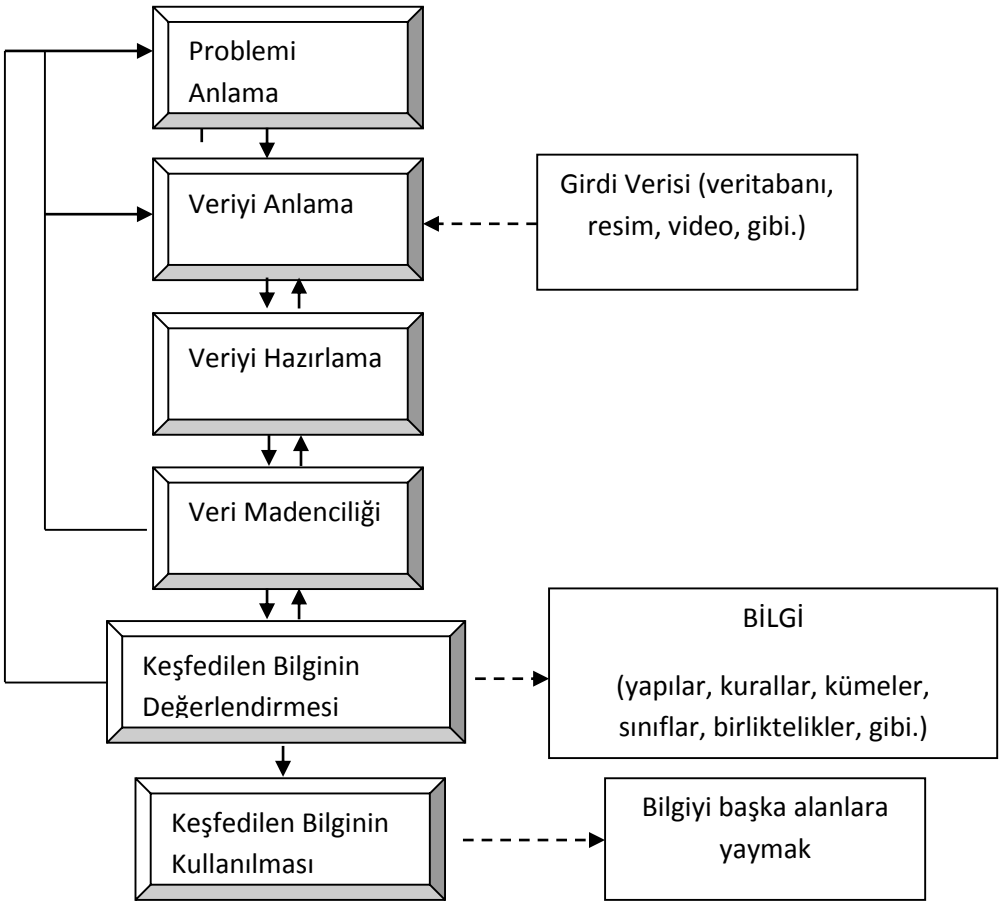
Sonuç olarak elde edilen, keşfedilen bilgi, müşterinin kullanacağı şekilde yönetilmeli ve sunulmalı.

- **Yayılım Planı:** Veri madenciliği sonuçlarını iş hayatına uygulamak için, değerlendirme sonuçları alınır ve yayılım için bir strateji oluşturulur.
- **Kontrol ve bakım planı:** Eğer veri madenciliği sonuçları işin ve onun çevresinin gündelik bir parçası olacaksa kontrol ve bakım önemli bir mesele olmaktadır.
- **Son raporun oluşumu:** Proje sonunda proje lideri ve ekibi nihai raporu hazırlarlar. Yayılım planına bağlı olduğu için, bu rapor bekli projenin ve onun deneyimlerinin bir özeti veya veri madenciliği sonuçlarının yansıtıldığı anlaşılır bir sunum olabilir.
- **Projenin Gözden Geçirilmesi:** Proje esnasında doğru ve yanlış gidenlerin değerlendirilmesi. İyi yapılanların neler olduğu ve geliştirilmesi gereken ihtiyaçların belirlenmesi. Kısaca proje deneyiminin özeti.

Kullanım alanları: Tıp, mühendislik, pazarlama ve satış alanlarında kullanılmaktadır.

2.3 Melez Modeller

Akademik ve endüstriyel modellerin gelişimi aynı zamanda melez modellerin de gelişimini sağladı. Cios vd. tarafından geliştirilen altı adım BKS modeli CRISP-DM modelinin akademik araştırmaya adaptasyonudur.



Şekil 2.3 Melez modellerin yapısı

Adım açıklamaları;

Problemi Anlama: Konuyla ilgili uzmanların projenin amacını belirleme ve problemin tanımı çalışmalarını içerir. Proje amacının VM amacına dönüştürülmesi ve VM aracının seçilmesi.

Veriyi anlama: Örnek verileri toplandığı ve format ve boyutta dahil olmak üzere hangi verilere ihtiyaç duyulduğunun belirlendiği adım. Veriler eksiklik, fazlalık, eksik değerler, değişken değerlerinin anlamlılığı gibi durumlar için kontrol edilir. Bu şekilde verilerin VM hedefine göre faydalı olup olmadıkları doğrulanmış olmaktadır.

Veri hazırlığı: İlerleyen adımda VM yöntemine girdi olarak kullanılacak veri yapısına karar verilen adımdır. Örnekleme, korelasyon ve anlamlılık testleri, veri temizleme, veri kayıtlarının eksiksizliğinin kontrolü, eksik, sapan değerlerin silinmesi veya düzeltilmesi gibi adımları içerir. Temizlenen veriler artık, yeni değişkenlerin oluşturulması, boyut indirgeme gibi süreçlere girebilir. Sonuç olarak, birinci adımda seçilen VM aracı için gerekli girdi ihtiyaçlarına sahip veri yapısı oluşturulur.

Veri Madenciliği: Hazırlanan veriden bilgi elde edebilmek adına birçok VM yöntemleri kullanılır.

Keşfedilen bilginin değerlendirilmesi: Bu adım, sonuçları anlama, keşfedilen bilginin yeni ve ilginç olup olmadığının tespiti, alan uzmanları tarafından sonuçların yorumu ve keşfedilen bilginin etkilerinin kontrolünü içermektedir.

Keşfedilen bilginin kullanımı: Son adımda keşfedilen bilginin nasıl kullanılacağı planlanır. Mevcut uygulama alanları genişletilebilir. Keşfedilen bilginin hayata geçirilmesinin kontrolü için plan oluşturulur ve tüm proje belgelendirilir. Sonunda keşfedilen bilgi yayılır.

Kullanım alanları: Tıp ve yazılım geliştirme alanlarında kullanılmaktadır.

2.3.1 SEMMA (Sample; Explore; Modify; Model; Assess)

CRISP-DM' e ek olarak, diğer bir veri madenciliği metadolojisi SAS enstitüsü tarafından Enterprise Miner veri madenciliği programı ile gerçekleşmektedir. Burada "S" örnekleme, bir yada daha fazla veri tabloları oluşturmaktır. Örnekler yeterince anlamlı bilgi içerecek kadar büyük ve işlenebilecek kadar küçük olmalıdır. "E" araştırma, veriyi anlama ve fikir elde edilme adına, beklenen ilişkiler, beklenmedik trendler ve anormalliklerin araştırılmasına gereksinim vardır. Modelleme işlemi öncesinde de verilerin İşlenmesi gerekmektedir. Bir takım değişken dönüşümleri ve değişken seçimleri bu aşamada gerçekleşir. Daha sonra analitik araçların ihtiyaç duyulduğu modelleme kısmına geçilir ve veriler içerisindeki yapıların belirlenir. Değerlendirme aşamasında da veri madenciliği işleminden elde edilen bulguların faydaları ve güvenilirliği ortaya çıkarılır. Büyük miktardaki veri yapılarına uygulanan SEMMA ile veri içerisinde bilinmeyen yapıların önceden ortaya çıkarılmasını sağlayarak önemli bir işletme avantajı sağlar. Bu süreç farklı iş problemlerinde uygulanabilmektedir; hilekarlık tespitleri- özellikle kredi kartı güvenlikleri, risk analizleri, müşteri memnuniyet analizleri, portföy analizleri, pazar segmentasyonu, müşteriyi elde tutma ve birliktelik analizleri gibi.

2.3.1.1 SEMMA Sürecinin Adımları

Adım 1 (örnekleme): Büyük veri tabanı içerisinden anlamlı ve yeterli veri kümesinin seçilmesi. Anlamlı bilgi içerecek kadar yeterince büyük, manipülasyon yapılabilecek kadar küçük bir örnek. Güvenilir ve tüm datayı temsil edebilecek örnekleme stratejisi,

maliyet ve hesaplama performansı gibi işlem zamanını azaltacaktır, böylece ticari olarak daha hızlı aksiyon alınabilecek ve bilgiye daha çabuk ulaşılacaktır.

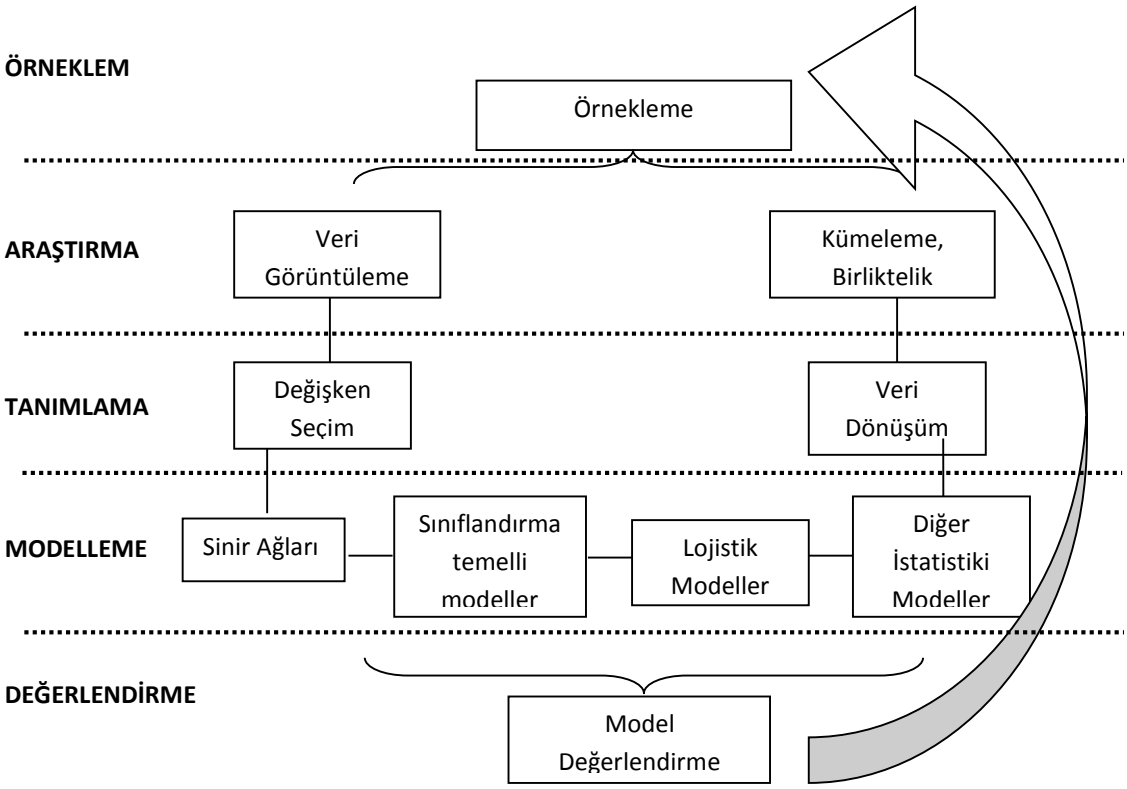
- Eğitim, uygun model bulmak için kullanılır
- Onaylama
- Test

Adım 2 (araştırma): Veri kümesini daha iyi anlayabilmek adına, veri çerisindeki anormallikleri ve beklenmeyen trendlerin incelenmesi, araştırılması. Örnekleme sonrası görsel ve sayısal olarak veri yapısal trendler ve kümeler için incelenir. Araştırma keşif sürecinin daha rahat işlemesini ve bu sürecin yönlendirmesinde yardımcı olmaktadır. Görsel araştırmada belirli trend yapıları ortaya çıkmıyorsa, data için faktör analizi, kümeleme analizi gibi istatistiksel teknikler kullanılabilmektedir.

Adım 3 (Tanımlama): Modelin inşası üzerine odaklanmak için, değişkenlerin oluşturulması, seçilmesi ve dönüşümleri gerçekleştirilir. Araştırma aşamasındaki keşiflere dayanılarak veriyi manipüle etmek gerekebilir. Yeni değişkenlerin türetilmesi, sapan değerlerin tespiti ve değişken sayısının azaltılması gibi. Veri madenciliği dinamik ve iteratif bir işlem olduğu için, yeni bir bilgiye ulaşıldığında model ve teknikler de güncellenebilir.

Adım 4 (Modelleme): Veri hazır olduktan sonra veri içerisindeki yapıları ortaya çıkaracak modelin inşası başlamaktadır. Veri madenciliğinde model teknikleri, yapay sinir ağları, karar ağaçları, lojistik modeller, zaman serisi analizleri, regresyon ve temel bileşenler analizi gibi istatistiksel yöntemleri içermektedir. Her bir model özel etki ve güce sahiptir ve her bir model veriye bağlı veri madenciliği konularına uygunluk gösterir. Örneğin, yapay sinir ağları lineer olmayan karmaşık ilişkilerde kullanılmaktadır.

Adım 5 (değerlendirme): Veri madenciliği sürecinde faydalı ve güvenilir bilgilerin ortaya çıkarılması. Bu son adımda gerçekleştirilen modellerin ne kadar iyi çalıştığını tahmin etmek için modeller değerlendirilir. Elde edilen model başka örnekler üzerinde test edilir. Eğer model doğruysa, model inşa edilirken kullanılan örnekte olduğu gibi, diğer örneklerde de çalışmalıdır. Eğer bir müşterinin elde tutma oranının yüksek olduğunu biliyorsak ve modelimizde elde tutma oranlarını tahmin edebiliyorsa, modelin bu müşteriyi seçip seçmediği kontrol edilebilir.



Şekil 2.4 SEMMA süreç tasarımı

BÖLÜM 3

KÜMELEME

“Dünya, onu oluşturan varlıklar arasındaki benzerlik ve farklılıklarının kavranılmasıyla anlaşılabilir” (Tyron ve Bailey, 1970).

Günümüzde bilgi zorunlu bir şekilde günlük yaşamın vazgeçilmezi olmuştur. Bu bilgi finans, medikal, bilimsel alanlarda olabileceği basit bir market için bile artık değerlidir. Özellikle internetin ve web tabanlı uygulamaların kullanımlarının yayılımı bilgi kullanımının da dünya genelinde yayılmasını sağlamıştır. Hem kullanıcı sayısının artması, hem de bilgi hacminin artması, bilgi yönetiminin ve verilerin daha karmaşık yapılara dönüşmesine sebep olmaktadır. Gerçek yaşamda uygulamalar yüksek boyutlu veriler oluşturmaktadır. Veriler limitsiz bir şekilde miktar ve boyut olarak artış göstermektedir. Artan, biriken, oluşan bu verilere rağmen yinede veriyi anlayabilmek, içerisinden anlamlı bilgiler elde etmek vazgeçilmezdir.

Doğası gereği veya örneklerle kazanılmış bazı bilgiler ile gözlemler, nesneler belirli sınıflara atanmaktadır. Sınıflama, verileri anlamaya ve anlamlı bilgiler elde etmekte oldukça yararlıdır. Fakat büyük sayıda gözlemlerden oluşan veri tabanlarına sınıf değerleri atayabilmek maliyetli bir işlem olacaktır. Diğer taraftan gözlemlerin atanacağı bir sınıf değeri, bu değere ilişkin ön bilgi gerçekten olmayabilir. Kümeleme verileri sınıf yada kümeler içinde gruplama sürecidir. Bu süreç sonunda küme içindeki gözlemler diğer kümelerdekine göre birbirine daha çok benzerlik gösteren, bir o kadar da diğer kümelerdekilerle farklılık gösteren grupların oluşması sağlanmaktadır. Kümelemenin kökleri veri madenciliği başta olmak üzere, istatistik, biyoloji ve makine öğrenimi alanlarında bulunmaktadır.

3.1 Kümeleme Tanımı

Fiziksel veya soyut nesnelerin, benzer nesnelerin bulunduğu sınıflar içerisinde gruplama süreci kümeleme olarak adlandırılır (Jiawei Han, 2006). Küme içerisindeki nesneler topluluğu aynı küme içerisindeki diğer nesneler ile benzer olup, diğer kümelerdeki nesneler ile farklılık göstermektedirler. İnsan doğası gereği çevresindeki nesneleri ve soyut olarak gerçekleşen olayları kategorileştirir, sınıflandırır veya gruplandırır ve hayatı da bu gruplamalar ve sınıflamalar sayesinde öğrenmektedir. Kümeleme analizi birçok uygulamada geniş kullanım alanına sahiptir bunlardan bazıları, pazar araştırmaları, yapı tanıma, veri analizleri, görüntü şekillendirme gibi. Perakende ve işletmelerde ise kümeleme, farklı müşteri gruplarını keşfetme, alışveriş alışkanlıkları ile müşteri gruplarının profillerinin oluşturulabilmesinde pazarlamacılara yardımcı olabilmektedir. Biyolojide benzer fonksiyon sergileyen genlere sahip olanların kategori edilmesi ve öz yapıları hakkında bilgi elde edilmesi gibi bitki ve hayvan sınıflandırmalarında kullanılmaktadır. Bunların yanında, bilgi keşfi için web de doküman sınıflandırılması, sapan değerlerin belirlenmesi, hane halkı veya bölge olarak ev tipi, değeri ve coğrafi konumu göz önünde bulundurularak benzer bölgeler tespit edilebilmektedir.

Sapan değerlerin tespiti aslında kredi kartı sahtekarlığı ve elektronik ticarette suçsal olayların takibinde de kullanılmaktadır. Örneğin kredi kartı işlemlerinde, müşterilerin harcama alışkanlıklarına göre (harcama sıklığı, ortalama harcama tutarları, ağırlıklı yapılan alışveriş türü vb.) kümeleme analizi sonucunda belirli gruplar elde edilir. Bu sonuçlar ışığında tanımlanacak belirli tolerans limitleri ile kurallar tanımlanabilir ve anlık (real-time) aksiyon alınabilir.

Makine öğreniminde, kümeleme denetimli olmayan öğrenim yöntemidir. Sınıflandırmanın aksine **denetimli olmayan öğrenim** (unsupervised learning) yöntemleri ve kümeleme önceden tanımlanmış sınıflara dayandırılmamaktadır. Bu nedenle kümeleme örneklerle değil, gözlemlerle öğrenen bir süreçtir. Kümelemenin amacı daha çok tanımlayıcı olmasına karşın, sınıflandırma daha çok tahmin amaçlıdır (Veyssieres ve Plant, 1998). Kümeleme algoritmalarının etkinliği ve yeterliliği özellikle geniş ve büyük veritabanına sahip işletmeler için daha değerli olmaktadır. Jiawei Han kümelemenin gereksinimlerine yönelik birkaç bilgi önermiştir:

Ölçeklendirilebilirlik: Mevcut birçok kümeleme algoritması küçük veri kümelerinde iyi çalışabilir, ancak uygulamada çoğunlukla büyük veri kümeleri ile karşılaşılır. Büyük veri kümelerinden seçilen örneklere uygulanan kümeleme yanlı sonuçlar oluşturabilir. Bu nedenle yüksek derecede ölçeklendirilebilir kümeleme algoritmalarına ihtiyaç bulunmaktadır.

Farklı Tip Değişkenlerin Varlığı: Birçok algoritma aralıklı ölçekli gözlemler içeren veri yapılarına uyum göstermektedir. Fakat uygulamalar ikili, sıralı, kategorik veya karışık veri tiplerinde olabilirler.

Keyfi küme Şekli Belirleme: Birçok kümeleme algoritması Öklidyen ve Manhattan mesafe ölçülerine dayanılarak tanımlanmıştır. Bu şekil mesafe ölçülerine dayandırılan algoritmalar benzer boyut ve yoğunlukta küresel şekil bulmaya eğilimlidir. Fakat küme herhangi bir şekil olabilir, yeter ki isteğe bağlı şekil kümeleri için algoritma geliştirilsin.

Girdi Parametrelerini Belirleyecek Bilgi için Asgari Gereksinimler: Analiz öncesi bir çok algoritma kullanıcıdan bazı parametreler ister(örneğin küme sayısı). Kümeleme sonuçları bu parametrelere çok duyarlıdır bu nedenle bu girdi parametreleri kümeleme sonuçlarına ciddi etkiler oluşturabilir. Özellikle yüksek boyutlu gözlemlerin bulunduğu veri kümelerinde, bu parametreleri kararlaştırmak çoğunlukla zordur.

Veri Olumsuzlukları: Gerçek uygulamadaki birçok ver kümelerinde sapan değerler, eksik, bilinmeyen veya hatalı gözlemler sıklıkla karşılaşılan bir durumdur. Bazı kümeleme algoritmaları bu durumlara karşı duyarlıdır ve bu tür veriler hatalı kümelenmelere yol açabilir.

İlave Kümeleme ve Yeni Kayıtlara Duyarsızlık: Bazı kümeleme algoritmaları veritabanı güncellemeleri sonucu oluşan yeni veri eklentilerini mevcut kümeleme yapıları içine entegre etmesi mümkün olamamaktadır. Bu nedenle yeni bir kümeleme işleminin başlaması gerekmektedir. Bazı algoritmalar ise girdi verilerinin sırasına duyarlıdır. Benzeri sorunları aşabilmek için ilave kümeleme algoritmaları ve girdi datasından etkilenmeyen algoritmaların geliştirilmesi gerekmektedir.

Veri Boyutunun Yüksek Olması: Veri ambarları ve veritabanı yüksek boyutlu ve bir çok değişken içerebilmektedir. Birçok algoritmalar düşük veri boyutlarında çok etkin

sonuçlar üretebilmektedir. Yüksek boyutlu veri uzayda kümeleme daha zor olmaktadır, çünkü veri daha dağınık ve çarpık olabilmektedir.

Kısıt Temelli Kümeleme: Gerçek uygulamalarda kümeleme süreçlerinde birçok kısıt olabilmektedir. Örneğin, yeni bir mağaza açabilmek için bölgesel kümeleme yaptığımızı düşünürsek, bazı kısıtlarımız var demektir; bölgeden bir nehir geçiyor olabilir, alt geçit, metro vb. kısıtlarla karşılaşılabilir.

Yorumlanabilirlik ve Kullanılabilirlik: Kümeleme sonuçları kullanıcı için kolay anlaşılır, yorumlanabilir ve kullanılabilir olmalıdır. Kümelemenin belirli, anlamlı yorumlara ve uygulamalara bağlı olmaya ihtiyacı olabilir.

3.2 Kümeleme Analizinde Veri Tipleri

3.2.1 Uzaklık (mesafe) Ölçüleri

Kümeleme birbirine benzer, yakın olan nesnelerin gruplaması olduğu için, iki nesnenin birbirine benzer yada farklı olduğuna dair bir ölçünün olması gerekmektedir. Bu ilişkiyi ortaya koyabilmek iki tip ölçüm gerçekleşmektedir: Benzerli ve Başkalık ölçütleri. Benzerlik ve başkalık ölçütleri kümeleme analizinin temel yapısını oluşturmaktadırlar. Veri kümesinde yer alan gözlemleri kıyaslayabilmek için, onların arasında bir mesafe ölçütü veya bir yakınlık bulmaya gereksinim vardır. Gözlem çiftleri arasındaki yakınlık indeksleri, homojen grupları belirleyebilmek için gerekli başlangıç bilgileri sağlar.$x_i$vex_j gözlemleri arasındaki mesafe $d(x_i, x_j)$ ile gösterilmektedir. Farklı değişken niteliklerine göre farklı başkalık ölçütleri kullanılabilmektedir.

3.2.1.1 Veri Matrisi

N gözlem, örneğin bir tüketici ve p değişken, yaş, boy, cinsiyet, ırk gibi, davranış veya ölçümler sonucu elde edilen bir veri matrisi şekildeki gibidir.

$$\begin{bmatrix} x_{11} & . & x_{1f} & . & x_{1p} \\ . & . & . & . & . \\ . x_{in} & . & x_{if} & . & x_{ip} \\ . & . & . & . & . \\ x_{n1} & . & x_{nf} & . & x_{np} \end{bmatrix}$$

3.2.1.2 Başkalık Matrisi

Bütün gözlem çiftlerinin yer aldığı ve yakınlık indekslerinden oluşan matris;

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ . & . & . & & \\ d(n,1) & d(n,2) & . & . & 0 \end{bmatrix}$$

Burada $d(i,j)$ i ve j gözlemi arasındaki farktır. Birçok kümeleme algoritması bu başkalık ölçütü üzerine kuruludur. Kümeleme analizi öncesinde veri matrisi üzerinden, başkalık matrisinin elde edilmesi ile başlanabilir.

$d(i,j)$ negatif olmayan bir sayıdır ve *i* ve *j* birbirine benzerse 0'a yakındır, birbirlerinden farklı ise daha büyüktür.

3.2.1.3 Aralıklı Ölçek Değişkenler

Aralıklı ölçek değişkenler genellikle doğrusal ölçekli sürekli ölçümlerdir. Ağırlık, uzunluk, koordinatlar, hava sıcaklığı gibi değerlerdir. Ölçüm birimi kümeleme analizini etkileyebilir. Örneğin harcama tutarını, TL yerine dolar olarak değiştirmek veya metre olan ölçüm birimini inç olarak değiştirmek farklı kümeleme yapılarının oluşmasına neden olabilir. Küçük ölçüm birimleri değişkenlerin daha geniş aralıklarda değerlere yönlendiriyor ve bu kümeleme yapısında daha büyük etkilerle neden olmaktadır. Ölçüm birimi seçiminin bu etkisinden kurtulmak için veri standardize hale getirilmelidir. Standardizasyon tüm verilere eşit ağırlıkların verilmesidir. Veri hakkında öncül bir bilgi olmadığı sürece bu faydalı bir yaklaşımdır. Fakat bazı uygulamalarda, bazı değişken kümelerine diğerlerine göre daha ağırlık verilebilmektedir. Örneğin bir basketbol adayı için boy uzunluğu değişkenine daha fazla önem atfedilebilir.

3.2.1.4 Mesafe Ölçüm Yöntemleri ve Standardizasyon

Standardizasyon, orijinal ölçümlerin birimsiz ölçümlere dönüştürülmesidir. f değişkeni için aşağıdaki yolla hesaplanır;

1. Ortalama Mutlak Sapma, s_f

$$s_f = \frac{1}{\mathrm{n}}(|\mathrm{x_{1f}} - \mathrm{m_f}| + |\mathrm{x_{2f}} - \mathrm{m_f}| + \cdots + |\mathrm{x_{nf}} - \mathrm{m_f}|), \qquad (3.1)$$

$\boldsymbol{x_{1f}, \cdots x_{nf}}$ f değişkeninin n ölçüsüdür.$\boldsymbol{m_f}$f' nin ortalama değeridir.

$$m_f = \frac{1}{n}\left(x_{1f} + x_{2f} + \cdots + x_{nf}\right), \tag{3.2}$$

2. Standardize ölçüm, z-skor

$$z_{if} = \frac{x_{if} - m_f}{s_f} \tag{3.3}$$

Ortalama mutlak sapma, s_f, standart sapmaya (σ_f) göre sapan değerlerde daha robust. Bu şekilde hesaplanan ortalama mutlak sapmada sapan değer etkisi düşer ($|\mathrm{x_{if}} - \mathrm{m_f}|$ karesi değil), diğer etkili dağılım ölçüsü ise "medyan mutlak sapma" dır. Fakat ortalama mutlak sapmayı kullanarak hesaplanan z-skorlarında sapan değerler küçük çıkmamaktadır ve denetimi kolaydır.

İhtiyaç duyuluyorsa standardizasyon gerçekleştikten sonra, ihtiyaç duyulmuyorsa ham verilerde aralıklı ölçekli gözlemler arasındaki benzerlik (yada başkalık), her gözlem çifti arasındaki mesafe ile hesaplanır. En çok kullanılanlarından "**Öklidyen Mesafesi**",

$$d(i,j) = \sqrt{\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 + \cdots + \left(x_{in} - x_{jn}\right)^2}, \tag{3.4}$$

$i = (x_{i1}, x_{i2}, \cdots, x_{in})vej = \left(x_{j1}, x_{j2}, \cdots, x_{jn}\right)$ iki n boyutlu gözlem verileridir.

Diğer "**Manhattan (city block) Mesafesi**",

$$d(i,j) = \left|x_{i1} - x_{j1}\right| + \left|x_{i2} - x_{j2}\right| + \cdots + \left|x_{in} - x_{jn}\right| \tag{3.5}$$

Her iki mesafe ölçüsü de şu matematiksel gereksinimleri sağlamaktadır,

1. $d(i,j) \geq 0$: mesafe negatif bir sayı değildir.
2. $d(i,i) = 0$: gözlemin kendisiyle mesafesi 0 dır.
3. $d(i,j) = d(j,i)$: mesafe simetrik bir fonksiyondur.
4. $\mathrm{d(i,j)} \leq \mathrm{d(i,h)} + \mathrm{d(h,j)}$: üçgen eşitsizliği.

$\mathrm{i} = \left(\mathrm{x_{i1}, x_{i2}, \cdots, x_{ip}}\right)$vej $= \left(\mathrm{x_{j1}, x_{j2}, \cdots, x_{jp}}\right)$olmak üzere iki p boyutlu gözlem verileridir. "**Minkowski Mesafesi**" hem Öklidyen hemde Manhattan mesafesinin genelleştirilmiş halidir. (Han ve Kamber, 2001):

$$\mathrm{d(i,j)} = \left(\left|\mathrm{x_{i1} - x_{j1}}\right|^{\mathrm{p}} + \left|\mathrm{x_{i2} - x_{j2}}\right|^{\mathrm{p}} + \cdots + \left|\mathrm{x_{in} - x_{jn}}\right|^{\mathrm{p}}\right)^{\frac{1}{\mathrm{p}}} \tag{3.6}$$

P pozitif tamsayıdır. Bu mesafe ölçüsü ayrıca L_p normu olarak da adlandırılır. Dolayısı ile p=1 için Manhattan, p=2 için Öklidyen mesafe ölçüleri, p= ∞ için Chebychev metriği elde edilmektedir.

Ölçüm birimi kümeleme analizine önemli derecede etki etmektedir. Bu ölçek birimi seçimine olan bağımlılıktan uzak durmak için veri standardize edilmelidir. Standardize edilmiş ölçümler tüm değişkenlerin eşit ağırlığa sahip olmaları ile gerçekleşmektedir. Fakat önem derecelerine göre her değişkene bir ağırlık atanmış olsaydı **Ağırlıklandırılmış Öklidyen Mesafe,**

$$d(i,j) = \sqrt{w_1|x_{i1} - x_{j1}|^2 + w_2|x_{i2} - x_{j2}|^2 + \cdots + w_n|x_{in} - x_{jn}|^2} \text{ ve } w_i \in [0,\infty) \quad (3.7)$$

Manhattan ve Minkowski mesafeleri de ağırlıklandırılmış değişkenlerle hesaplanabilir.

3.2.2 İkili Değişkenler için Mesafe Ölçütleri

İkili değişkenler sadece iki duruma sahiptir, 0 veya 1, 0 değişken boş, 1 ise değişken dolu demektir. Örneğin bir kişinin iş durumunu tanımlıyorken, 0 "çalışmıyor", 1 ise "çalışıyor" u ifade etmektedir. İkili değişken yapısına, aralıklı ölçekmiş gibi davranmak yanlış kümeleme sonuçlarına yol açmaktadır. Bu nedenle farklı bir benzerlik (başkalık) ölçütünün kullanılması gerekmektedir. İkili değişkenler için çapraz tablolara dayanarak mesafe ölçütleri hesaplanabilmektedir. Çizelge 3.1' de 2x2 çapraz tablosu yer almaktadır. Değişkenlerin toplam sayısı $p = q + r + s + t$. İkili değişkenler her iki durumda da eşit değere ve ağırlığa sahiplerse yani sonuçların hangisinin 1, hangisinin 0 ile kodlanacağı tercihi yoksa simetrik ikili değişkenlerdir. Örneğin cinsiyet, "erkek" ve "bayan". Simetrik ikili değişkenlerde başkalık, simetrik ikili başkalık olarak adlandırılır. (2.8)' deki eşitlik *i* ve *j* gözlemleri arasındaki başkalık ölçüsüdür.

$$d(i,j) = \frac{r+s}{q+r+s+t} \quad (3.8)$$

Burada q her iki gözlemde 1' e eşit niteliklerin toplam sayısıdır; t her iki gözlem için 0 değerine sahip niteliklerin toplamıdır; *r* ve *s* her iki gözlem için eşit olmayan durumların toplamıdır.

İkili değişkenlerde durumların sonuçları eşit anlamda önemli değilse asimetrik ikili değişkenlerdir. Örneğin bir hastalık sonuçlarındaki pozitif veya negatif durumlar (genellikle pozitif durumlar önemlidir). Bizim için önemli bir sonucu veya çok nadir bir

durumu 1 olarak, diğer durumun sonucunu da 0 olarak kodlayabiliriz. Bu durumda iki 1' in eşleştiği göze; iki 0' ın eşleştiği gözeye göre daha anlamlıdır. Bu şekildeki değişkenlere dayanan başkalık asimetrik ikili başkalık olarak adlandırılmaktadır. Bu nedenle t(negatif eşleşmeler) önemsiz durum olarak düşünülür ve (3.9)' daki eşitlikte olduğu gibi hesaplamadan (paydadan)çıkarılır. Jaccard coefficient olarak adlandırılmaktadır.

$$d(i,j) = \frac{r+s}{q+r+s} \tag{3.9}$$

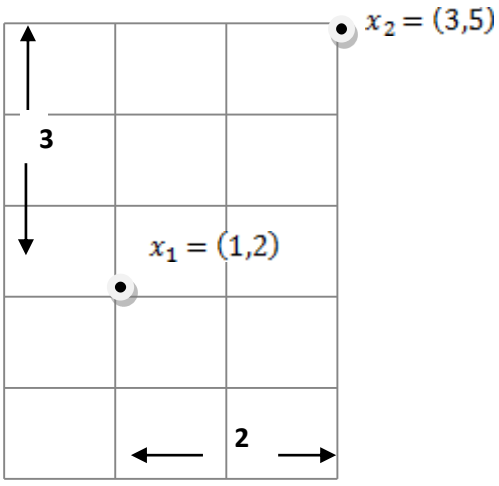
Tümleyeni olarak, iki ikili değişken arasındaki mesafenin benzerlik ölçütü hesaplanabilir. i ve j arasındaki asimetri ikili benzerlik ölçütü, $sim(i,j)$;

$$sim(i,j) = \frac{q}{q+r+s} = 1 - d(i,j) \tag{3.10}$$

$sim(i,j)$' nin katsayısı jaccard katsayısı olarak adlandırılmaktadır.

Çizelge 3.1 İkili değişkenler için çapraz tablo

		Gözlem j		
		1	0	toplam
Gözlem i	1	q	r	q+r
	0	s	t	s+t
	Toplam	q+s	r+t	



Öklidyen mesafe

$(2^2 + 3^2)^{\frac{1}{2}} = 3.61$

Manhattan mesafe

$2 + 3 = 5$

Şekil 3.1 İkili değişkenler mesafe ölçüsü hesaplama örneği

3.2.3 Kategorik Değişkenler için Mesafe Ölçütleri

Kategorik değişkenler ikili değişkenlerin genelleştirilmiş hali olup ikiden fazla durumun olduğu değişkenlerdir. Bu tür değişkenler her ne kadar bir tamsayı ile ifade edilseler de, sembol veya harflerle de ifade edilebilirler. Dolayısı ile atanan sayılarda herhangi bir tanımlayıcı sıralama yoktur. Örneğin beyaz, siyah, kırmızı, mavi, yeşil' den oluşan 5 renk kategorik değişkenler olabilir. *i* ve *j* arasındaki başkalık ölçütü yanlış eşleşme oranıyla hesaplanabilmektedir:

$$d(i,j) = \frac{p-m}{p}, \tag{3.11}$$

m eşleşme sayısı (*i* ve *j*' de aynı durumdaki değişkenlerin sayısı), *p* toplam değişken sayısıdır.

Çizelge 3.2 Değişken tiplerine göre 4 gözleme ait örnek veriler

Gözlem	**Test1-Kategorik**	**Test2-Sıralı**	**Test3-oransal**
1	A	Çok iyi	445
2	B	Ortalama	22
3	C	İyi	164
4	A	Çok iyi	1210

Çizelge 3.2' ye göre başkalık matrisi,

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix}$$

kategorik değişken test-1, eşitlikteki *p*=1' dir. $\mathrm{d(i,j)}$*i* ve *j* eşleşirse 0, farklıysa 1 değerini alacaktır,

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Şeklinde başkalık matrisi elde edilebilmektedir.

3.2.4 Sıralı Değişkenler için Mesafe Ölçütleri

Kesikli sıralı değişkenleri kategorik değişkenlere benzerlik göstermektedir. Fakat bu sefer değişkenlere ait durumların aldıkları değerler anlamlı bir sıra izlemektedir. Sıralı değişkenler objektif olarak ölçülemeyen ölçümlerin durumları hakkında kıyaslama yapabilecek yararlı bilgiler sunabilmektedir. Örneğin bir şirkette asistan, yönetici, müdür, genel müdür ardışık sıralı ifade edilebilmektedir. Sürekli sıralı değişkenler ise bilinmeyen bir ölçekteki sürekli küme gibidir. Örnek olarak spor müsabakalarındaki bronz, gümüş, altın.

Sıralı değişkenin *f* olduğunu farz edelim ve bunun M_f durumu olsun. Dolayısıyla durum sıralaması $1, \dots, M_f$ olacaktır. Sıralı değişkenlerde hesaplanan gözlemler arasındaki başkalık hesabı aralıklı ölçek değişkenlerindekine benzerdir. n gözlem bulunan f sıralı değişken kümesi olduğunu düşünürsek. Başkalık ölçütü şu aşamaları izlemektedir,

1. f' nin i. gözlemi x_{if} ve f' nin M_f sıralı durumu olsun. Her x_{if} yi derecelerine ilişkin $r_{if} \in \{1, \dots, M_f\}$ yer değiştirilir.
2. Her bir sıralı değişken farklı durum sayılarına sahip olabileceği için, eşit ağırlığa sahip olabilmeleri için tüm değişken dizisini $[0,1]$ üzerine yansıtmak gerekebilir. f. değişkende i. gözlemin r_{if} sırasını
3.

$$z_{if} = \frac{r_{if}-1}{M_f-1} \quad (3.12)$$

4. Değiştirilerek standardize değerler elde edilmektedir.

5. Başkalık aralıklı ölçek değişkenlerinde kullanılan herhangi bir mesafe ölçütleriyle, i. gözlemin f değerini ifade edenz_{if} değeri kullanılarak hesaplanabilmektedir.

Çizelge 3.2' de yeralan örnek verilere üzerinden, gözlem değerleri ve test-2 de yeralan sürekli sıralı değişkende $M_f = 3$ durum bulunmaktadır. Önem sırasına göre 3, 1, 2 ve 3 olarak 4 gözleme değer atanır. Sonrasında bu değerler 0 ile 1 arasında bu değerler normalleştirilir; 1→0 olarak, 2→0.5 ve 3→1 olarak. Öklidyen Mesafe ile başkalık matrisi:

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1 & 0.5 & 0 \end{bmatrix}$$

3.2.5 Oran Ölçekli Değişkenler için Mesafe Ölçütleri

Oran ölçekli değişkenler doğrusal olmayan ölçeklerde pozitif bir ölçüm sağlamaktadır. Eksponansiyel ölçeklerde: Ae^{Bt}yadaAe^{-Bt}

A ve B sabit ve t de zamanı göstermek üzere, bir bakteri popülasyonunun büyümesi örneği verilebilir. Bu tarz ölçümler üç yöntemle hesaplanabilir.

6. Oran ölçekli değişkenlere aralıklı ölçek değişkenlermiş gibi hesaplamaya gidilebilir. Ölçeğin deforme olabilmesi nedeniyle genellikle iyi bir tercih olmamaktadır.

7. Logaritmik dönüşüm: oran ölçekli f değişkeninin i gözleminin x_{if} değeri $y_{if} = \log(x_{if})$ şeklinde dönüştürülür. y_{if} değeriaralıklı ölçekmiş gibi mesafe ölçütleri kullanılabilmektedir. Bazı oran ölçekli değişkenlere, değişkenin tanımı ve uygulamaya göre log-log veya diğer dönüşümler uygulanabilir.

8. x_{if} değerine sürekli sıralı değişkenmiş gibi uyarlayıp, sıra dizisine aralıklı ölçekmiş gibi davranıp mesafe ölçütleri hesaplanabilmektedir.

Örnek 7: Aynı tablodaki gözlem ve oransal ölçekli değişken olan test-3 ile logaritmik dönüşüm 1. gözlemden, 4. gözleme sırasıyla 2.65, 1.34, 2.21 ve 3.08. Öklidyen mesafe ölçütü ile başkalık matrisi,

$$\begin{bmatrix} 0 & & & \\ 1.31 & 0 & & \\ 0.44 & 0.87 & 0 & \\ 0.43 & 1.74 & 0.87 & 0 \end{bmatrix}$$

3.2.6 Karışık Tipli Değişkenler için Mesafe Ölçütleri

Aynı tipteki değişkenler arasındaki başkalık ölçümleri yukarıdaki işlemlerle gerçekleşebilmektedir. Ancak uygulamada yukarda bahsedilen tüm değişken tiplerini içeren gerçek veritabanları söz konusudur.

Karışık tipli değişkenler arasındaki başkalık ölçümü için bir yaklaşım aynı tür değişkenleri gruplayarak, her bir değişken tipi için ayrı kümeleme analizi çalıştırılır. Eğer uygun sonuçlar alınıyorsa bu mümkün olabilmekte. Ancak gerçek uygulamalarda, her bir değişken tipine farklı kümeleme analizi uygulamak uygun sonuçlar vermesi muhtemel değildir. Tüm anlamlı değişkenlerin $[0,1]$ arasındaki ortak ölçekte yansıtılıp tek bir başkalık matrisinin oluşumu sağlanabilmektedir. Daha çok tercih edilen tek bir kümeleme analizi ile tüm değişkenlerin beraber işleme girmesidir.

Karışık tipli *p* değişkeni içeren bir veri kümesinde *i* ve *j* gözlemleri arasındaki başkalık $d(i,j)$:

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}} \tag{3.13}$$

olarak tanımlanmaktadır. Gösterge:

$$\delta_{ij}^{(f)} = \begin{cases} 0, \ x_{if} \text{ yada } x_{jf} \text{ boş değerler } alması\ yada\ x_{jf} = x_{if} = 0 \\ \qquad\qquad ve\ f\ değişkeni\ asimetrik\ ikili\ değişken \\ 1, \ diğer\ durumlar \end{cases} \tag{3.14}$$

- i ve j gözlemlerinin arasındaki başkalığa f değişkeninin katkısı $d_{ij}^{(f)}$ değişken tipine göre hesaplanmaktadır.
- Eğer f aralıklı ölçek ise: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$,h, f değişkeninin tüm boş olmayan gözlemler için
- Eğer ikli veya kategorik değişken ise: $d_{ij}^{(f)} = \begin{cases} 0, & x_{if} = x_{jf} \\ 1, & değilse \end{cases}$

- Sıralı ise: r_{if}sıra değerleri vez_{if} standart değerler hesaplandıktan sonraz_{if} standart değerleri aralıklı ölçekmiş gibi$d_{ij}^{(f)}$ hesaplanır
- Eğer f oran ölçekli ise: $d_{ij}^{(f)}$logaritmik dönüşüm gerçekleştirerek aralıklı ölçek hesaplamalarının yapılması yada sürekli sıralı değişkenmiş gibi r_{if} ve z_{if}değerlerinin hesaplanması ve z_{if} aralıklı ölçek olarak düşünerek hesaplanır

Çizelge 3.2' de yeralan tüm değişkenleri göz önüne alındığında başkalık matrisi, test-1 ve test-2 için başkalık matrisleri ayrı ayrı hesaplanmıştı. Dolayısı ile eşitlikte yerlerine konması yeterlidir. Test-3 oransal ölçek için bir işlem gerekmekte. 2.65, 1.34, 2.21 ve 3.08 olarak logaritmik dönüşüm gerçekleştirilen değerlerin normalleştirilmesi gerekmektedir.$\max_h x_{hf} = 3.08$, $\min_h x_{hf} = 1.34$örnek 7 deki başkalık matrisinin her bir değerini (3.08-1.34)=1.74' e bölersek başkalık matrisi,

$$\begin{bmatrix} 0 & & & \\ 0.75 & 0 & & \\ 0.25 & 0.50 & 0 & \\ 0.25 & 1.00 & 0.50 & 0 \end{bmatrix}$$

Üç değişken için başkalık ölçütü xx deki eşitlikteki gibi hesaplanabilir, örneğin $d(2,1) = \frac{1(1)+1(1)+1(0.75)}{3} = 0.92$. 3 değişkenle tanımlanan veri için başkalık matrisi:

$$\begin{bmatrix} 0 & & & \\ 0.92 & 0 & & \\ 0.58 & 0.67 & 0 & \\ 0.08 & 1.00 & 0.67 & 0 \end{bmatrix}$$

olarak hesaplanır.

3.2.7 Vektör Gözlemler

Bazı uygulamalarda karmaşık gözlemlerin (şifre, ibare) bulunduğu sembolik kayıtların karşılaştırılması ve kümelenmesi gerekebilir, özellikle bilgi düzeltme, metin madenciliği, biyolojik sınıflandırmalarda, tf-idf (term frequency–inverse document frequency) belge vektörleri gibi veri yapılarında geleneksel metrik mesafe ölçüleri yerine metrik olmayan benzerlik fonksiyonları ile karşılaşılmaktadır.

$\boldsymbol{x}$ ve$\boldsymbol{y}$olan iki vektörü karşılaştırabilmek ve benzerlik fonksiyonu $s(\boldsymbol{x}, \boldsymbol{y})$tanımlayabilmek için birkaç yol bulunmaktadır. Bunlardan biri de **kosinüs(cosine) ölçüsü**dür:

$$s(x, y) = \frac{x^t.y}{\|x\|\|y\|} \quad (3.15)$$

$\boldsymbol{x}^t$, $\boldsymbol{x}$ vektörünün transpozudur, $\|\boldsymbol{x}\|$ ise $\boldsymbol{x}$ ' in öklidyen normu, $\|\boldsymbol{y}\|$ ise $\boldsymbol{y}$' nin öklidyen normudur, s ise $\boldsymbol{x}$ ve $\boldsymbol{y}$ olan iki vektörleri arasındaki açının kosinüsüdür. Bu değer rotasyon ve genleşmeye karşı değişim göstermez, fakat doğrusal dönüşümlerde değişim göstermektedir.

Eğer değişkenler ikili ise (0 yada 1)

Örnek $\boldsymbol{x} = (1, 1, 0, 0)$ ve $\boldsymbol{y} = (0, 1, 1, 0)$ olmak üzere iki vektör olsun, xx deki eşitlikten $\boldsymbol{x}$ ve $\boldsymbol{y}$ arasındaki benzerlik$s(\boldsymbol{x}, \boldsymbol{y}) = \frac{0+1+0+0}{\sqrt{2}\sqrt{2}} = 0.5$

3.2.7.1 Pearson Korelasyon Ölçüsü

$$s(x, y) = \frac{(x-\bar{x})^T(y-\bar{y})}{\|x-\bar{x}\|\|(y-\bar{y})\|} \quad (3.16)$$

$\overline{\boldsymbol{x}}$ tüm boyutlarda $\boldsymbol{x}$' in ortalamalarını göstermektedir.

3.2.7.2 Genişletilmiş Jaccard Ölçüsü (Tanimoto Katsayısı)

$$s(x, y) = \frac{x^t.y}{\|x\|^2+\|y\|^2-x^t.y} \text{ (Strehl ve Ghosh, 2000)} \quad (3.17)$$

3.2.7.3 Dice Katsayı Ölçüsü

$$s(x, y) = \frac{2x^t.y}{\|x\|^2+\|y\|^2} \quad (3.18)$$

Sonuç olarak Kümeleme Analizi için birçok benzerlik (başkalık) fonksiyonu ve verileri normalleştirme yöntemleri olmasına karşın, seçim yapılmasını gerektiren genel bir standart bulunmamaktadır. Ölçüm seçimlerinin uygunluğu direk olarak uygulamaya bağlı olacaktır. Mevcut uygulamada elde edilen kümelerin yararlı ve anlamlı olabilmesinden emin olabilmek için bu ölçüm tiplerine önem gösterilmelidir.

3.2.7.4 Değerlendirme Kriter Ölçüleri

Bir kümeleme işleminin iyi veya kötü olduğuna dair kesin bir değerlendirme yapmak sorun ve tartışmaya açık bir meseledir. Buna karşın birçok değerlendirme kriteri

literatürde geliştirilmiştir. Bu değerlendirme kriterlerini iki kategoride incelenmektedir: İçsel ve dışsal

3.2.7.5 İçsel Nitelik Kriteri

İçsel nitelik metrikleri belirli benzerli ölçütleri ile kümelerin kompakt bir yapıda olup olmadıklarını ölçmektedir. Bu genellikle küme içi homejenliği ve kümeler arası ayrılabilirliği yada bu ikisinin kombinasyonunun ölçümüdür. Verinin yanında ekstra bir bilgi gereksinimine ihtiyaç yoktur.

3.2.7.6 Hata Kareler Toplamı (SSE)

Kümeleme için oldukça geniş kullanıma sahip kriter ölçüsüdür:

$$SSE = \sum_{k=1}^{K} \sum_{\forall x_i \in C_k} \|x_i - \mu_k\|^2 \quad (3.19)$$

C_k, k kümesindeki örnek set; μ_k ise k kümesinin vektör ortalaması. μ_k' nın bileşeni:

$$\mu_{k,j} = \frac{1}{N_k} \sum_{\forall x_i \in C_k} x_{i,j} \quad (3.20)$$

$N_k = |C_k|$k kümesine ait örnek sayısı.

SSE kriterini minimize eden kümeleme yöntemi, minimum varyans bölümlendirmeleri olarak adlandırılmaktadır. Cebirsel manipülasyon ile,

$$SSE = \frac{1}{2} \sum_{k=1}^{K} N_k \bar{S}_k \quad (3.21)$$

$$\bar{S}_k = \frac{1}{{N_k}^2} \sum_{x_i, x_j \in C_k} \left\| x_i - x_j \right\|^2, \quad (3.22)$$

$(C_k = k\ k\ddot{u}mesi)$ olmak üzere.

SSE kriter fonksiyonu, birbirinden çok iyi ayrılmış kompakt bulut formundaki durumlar için uygun bir kriterdir (Duda vd., 2001).

3.2.7.7 Diğer Minimum Varyans Kriteri

SSE' ye ek minimum kriteri $\bar{S}_k$ değerini aşağıdaki ifade ile değiştirerek elde edilmektedir:

$$\bar{S}_k = \frac{1}{{N_k}^2} \sum_{x_i, x_j \in C_k} s(x_i, x_j) \text{ yada } \bar{S}_k = \min_{x_i, x_j \in C_k} s(x_i, x_j) \quad (3.23)$$

3.2.7.8 Serpilme Kriteri

Serpilme kriteri küme içi ve kümeler arası dağılımın ve her ikisinin toplam dağılım matrisini yansıtan serpilme matrislerinden elde edilmektedir. *k*. küme için serpilme matrisi:

$$S_k = \sum_{x \in C_k} (x - \mu_k)\,(x - \mu_k)^T \tag{3.24}$$

Küme içi serpilme matrisi:

$$S_W = \sum_{k=1}^{K} S_k \tag{3.25}$$

Kümeler arası serpilme matrisi:

$$S_B = \sum_{k=1}^{K} N_k (\mu_k - \mu)(\mu_k - \mu)^T \tag{3.26}$$

μgenelortalama vektör olup:

$$\mu = \frac{1}{m} \sum_{k=1}^{K} N_k \mu_k \tag{3.27}$$

Genel serpilme matrisi:

$$S_T = \sum_{x \in C_1, C_2, \dots, C_k} (x - \mu)(x - \mu)^T \tag{3.28}$$

Bu üç S_T, S_B, $S_W$serpilme matrisinden üç skalar kriter çıkabilir:

3.2.7.9 İz Matris Kriteri

Bir matristeki diyagonaldaki elemanların toplamı. S_W matrisinin izini minimize etmek SSE' nin minimizasyonuna benzerdir. Bu kriter küme içi serpilmeye dayanarak:

$$J_e = tr[S_W] = \sum_{k=1}^{K} \sum_{x \in C_k} \|x - \mu_k\|^2 \tag{3.29}$$

Diğer bir kriter ise kümeler arasını maksimize eden kriter olabilir:

$$tr[S_B] = \sum_{k=1}^{K} N_k \|\mu_k - \mu\|^2 \tag{3.30}$$

3.2.7.10 Determinant Kriteri

Serpilme matrisinin determinantı kabaca serpilme hacminin karesidir. Küme sayısı boyuttan küçük veya eşitse yada m-c değeri boyuttan küçükse S_B tekil matris olacağı için determinant uygun bir kriter olmayacaktır.S_W matrisinin tekil olmayan bir matris olduğu varsayımında determinant kriter fonksiyonu kullanılabilmektedir.

$$J_d = |S_W| = \sum_{k=1}^{K} S_k \tag{3.31}$$

3.2.7.11 Değişmezlik Kriteri

$S_W^{-1}S_B$ matrisinin öz değerleri olan $\lambda_1, \lambda_2, \dots, \lambda_d$ serpilme matrisinin temel doğrusal değişmezleridir. Öz değerlerden elde edilen birkaç kriter bulunmaktadır:

$$tr[S_W^{-1}S_B] = \sum_{i=1}^{d} \lambda_i \tag{3.32}$$

$$J_f = tr[S_T^{-1}S_W] = \sum_{i=1}^{d} \frac{1}{1+\lambda_i} \tag{3.33}$$

$$\frac{|S_W|}{|S_T|} = \prod_{i=1}^{d} \frac{1}{1+\lambda_i} \tag{3.34}$$

3.2.7.12 Condorcet Kriteri

$$\sum_{C_i \in C} \sum_{\substack{x_j, x_k \in C_i \\ x_j \neq x_k}} s(x_j, x_k) + \sum_{C_i \in C} \sum_{x_j \in C_i;\, x_k \notin C_i} d(x_j, x_k) \tag{3.35}$$

$s(x_j, x_k)$ ve $d(x_j, x_k)$ ölçümleri, x_j ve x_k vektörlerinin benzerlik ve başkalık ölçüleridir.

3.2.7.13 C-Kriteri

C-Kriteri (Fortier ve Solomon, 1996) **Condorcet Kriteri**

$$\sum_{C_i \in C} \sum_{\substack{x_j, x_k \in C_i \\ x_j \neq x_k}} \big(s(x_j, x_k) - \gamma\big) + \sum_{C_i \in C} \sum_{x_j \in C_i;\, x_k \notin C_i} \Big(\gamma - s\big(x_j, x_k\big)\Big) \tag{3.36}$$

γ eşik değerini göstermektedir.

3.2.7.14 Kategori Fayda Metriği

Kategori fayda (Gluck ve Corter, 1985) güvenilir bir kümeleme ile doğru bir şekilde tahmin edilen nitelik değerlerinin beklenen değerlerinin artışını tanımlanmaktadır. Bu metrik göreli az sayı içeren nominal değişkenlerin problemini içeren durumlar için faydalıdır.

3.2.7.15 Sınır Kesim Metriği

Bazı durumlarda kümeleme problemini, sınır kesim problemi olarak tasvir etmek faydalı olabilmektedir. Bazı örneklerde nitelik kalan sınır ağırlıklarının tüm sınır ağırlıklarına oranı olarak ölçülebilmektedir. Eğer kümelerin boyutları için bir kısıt yok ise, en iyi değeri bulmak çok kolaydır. Bu nedenle minimum kesim ölçüsü dengesiz yapıların kontrolü için revize edilir.

3.2.7.16 Dışsal Nitelik Kriterleri

Dışsal ölçümler küme yapılarının önceden belirlenmiş sınıflarla eşleşip eşleşmediğinin araştırılmasında faydalı olabilmektedir.

3.2.7.17 Ortak Bilgi Temelli Ölçüm

Ortak bilgi kriteri kümeleme için dışsal bir ölçü olarak kullanılabilmektedir. (Strehl vd., 2000).$C = \{C_1, \dots, C_g\}$ kullanılarak kümelenmiş ve tanım kümesi $dom(y) = \{c_1, \dots, c_k\}$ olan *y* değişkenine atfedilmiş *m* örnekleri için ölçüm:

$$C = \frac{2}{m} \sum_{l=1}^{g} \sum_{h=1}^{k} m_{l,h} \log_{g.k} \left(\frac{m_{l,h}.m}{m_{.,l}.m_{l,.}} \right) \quad (3.37)$$

$m_{l,h}$, C_l kümesindeki ve c_h sınıfındaki gözlem sayısını göstermektedir. $m_{.,h}$ ise c_h sınıfındaki toplam gözlem sayısını ifade etmektedir. Benzer şekilde $m_{l,.}C_l$ kümesindeki gözlem sayısıdır.

3.2.7.18 Kesinlik-Geri Çağırma Ölçüsü

Bilgi düzeltmeden gelen bu ölçü kümelerin değerlendirilmesinde dışsal ölçü olarak kullanılmaktadır. Küme belirli bir sınıf için bir sorgunun sonucu olarak incelenir. Kesinlik örnekten doğru bir şekilde alınan parçadır. Geri çağırma ise tüm eşleşen örneklerin dışında örneklerden doğru bir şekilde alınan kısımdır. Birleştirilen *F* ölçüsü kümeleme yapılarının değerlendirilmesi için faydalıdır (Larsen ve Aone, 1999).

3.2.7.19 Rand İndeks

Rand indeksi iki kümeleme yapısının karşılaştırılmasında kullanılmaktadır (Rand, 1971). C_1 ve C_2 olmak üzere iki küme yapısının olduğu düşünülürse, C_1 ve C_2 aynı kümeye atanmış *a* gözlem çifti sayısı; *b,* C_1 de aynı kümede olan, fakat C_2' de aynı kümeye atanmamış gözlem çift sayısı; *c*, C_2' de aynı kümeye atanmış, fakat C_1' de aynı kümeye atanmamış gözlem çift sayısı; *d*, C_1 ve C_2' de farklı kümelere atanan gözlem çift sayısı olsun. *a* ve *d* nicelikleri uyuşmaları, *b* ve *c* ise uyuşmazlıkları gösterdiği söylenebilir. Buna göre Rand İndeks:

$$RAND = \frac{a+d}{a+b+c+d} \quad (3.38)$$

Rand indeksi 0 ile 1 arasında değerler almaktadır. İki bölüm çok iyi uyuşmaktaysa rand indeksi 1' e yakın değerde olacaktır.

BÖLÜM 4

ALGORİTMA ve TEKNİKLER

4.1 Kümeleme Algoritmaları

Kümeleme yöntemlerinde kesin bir kategori yapısı olmamasına karşın Farley ve

Raftery (1998) kümeleme yöntemlerini iki ana grupta toplamaktadır: hiyerarşik ve bölümlendirme yöntemleri. Han ve Kamber (2001) ise bunlara ek olarak: yoğunluk temelli kümeleme, ızgara temelli kümeleme ve model temelli kümeleme yöntemlerini önermektedir.

1. Bölümlendirmeye dayalı kümeleme, hedef fonksiyona dayalı kümelemede denmektedir.
2. Sıradüzensel-hiyerarşik kümeleme
3. Yoğunluk temelli kümeleme
4. Izgara temelli kümeleme
5. Model tabanlı kümeleme

Bölümlendirmeye dayalı kümelemede, hedef fonksiyonunun minimizasyonu datanın içerisinde var olan yapının keşfedilmesini sağlar. Burada optimizasyon problemi çok iyi irdelenmeli. Bu ve benzeri yöntemlerde, küme sayıları önceden tanımlanır, hedef fonksiyonun minimizasyonu süreci işler.

Hiyerarşik kümelemenin temelinde ardıl gelişen kümelerin oluşumu vardır. Bu ya ardıl bölünmeler sonucu gerçekleşir, ki bu tek bir kümeden başlamaktır o da bütün verilerdir

yada data içerisinde yeralan her bir elemanın bireysel başlangıç kümesinin olmasıdır, ki burada da her iterasyonda birleşerek yeni kümeler oluşmaktadır

4.2 Hedef Fonksiyon (Bölümlendirme Kriteri) Temelli Kümeleme

Bölümlendirme algoritması, d-boyutlu gözlem uzayında *N* gözlemli *V* veri setinde $c(c \leq N)$sayıda küme formunun oluşmasını sağlamaktadır. Bu kümeleme yöntemi bazı performans indekslerinin (Hedef Fonksiyon) minimizasyonu yoluyla veri seti içerisinde gizli kalmış anlamlı yapıları araştırır. Doğru seçilmiş bir hedef fonksiyon ve onun minimizasyonu ile verideki gerçek yapının ortaya çıkmasında yardımcı olabilmektedir. Hedef fonksiyonu formüle etme ve eniyileme kurgularını organize etmenin birçok yolu bulunmaktadır. Yapının ortaya çıkarılmasında iki etken vardır; temsili küme ortalama vektörleri (prototipler) ve bölme matrisidir.

Prototipleri $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_c$ olarak gösterelim. Bölme matriside $U = [u_{ik}]$ ardışık kümelere ilgili verinin dağılımını tanımlayan girdileri içeren *c* sıra ve *N* kolon sayısından oluşan matristir.

4.2.1 K-Ortalamalar

K-ortalamalar Lloyd (1957, 1982), Forgey (1965), Friedman ve Rubin (1967), MacQueen (1967) algoritması kullanımı oldukça geniş bir kümeleme algoritmasıdır. Büyük data setlerinde veriyi kümeleme ve segmentasyonun (bölümlendirme) amacı doğal grupları belli olmayan verileri, birimleri veya değişkenleri anlamlı gruplara veya kümelere ayırabilmektir. Elde edilen bu küme veya gruplar **diğerlerine göre** kendi içlerinde daha benzer özelliklere sahip olan birimlerin veya nesnelerin yer aldığı grupları teşkil etmektedir. Regresyon ve sınıflandırma gibi denetimli yöntemlerde (supervised learning) sınıf etiketinin ve hedef değerin bilgisinin olmasına karşın, kümeleme prosedürüne girdi oluşturan nesnelerle ilişkili bir hedefle gelmemektedir. Bu nedenle kümeleme denetimsiz öğrenme yöntemi olarak atfedilir. Verilerin doğal grup bilgisine ihtiyaç duyulmadığı için, doğal grup etiket bilgisine ulaşmanın zor olduğu durumlarda denetimsiz öğrenme algoritmaları birçok uygulama için çok uygundur. Kümeleme gibi denetimsiz yöntemler, denetimli yöntemler kullanılmadan önce veri kümesini anlamada ve karakterize etmekte oldukça sık kullanılmaktadır.

Kümeleme sınıf etiketlerinden yararlanmasa da, nesnelerin davranışlarına yönelik bazı benzerlik kavramları tanımlanmalıdır. Nesnelerin kümelendiği benzerlik tanımı ve yöntemi uygulanan kümeleme algoritmasına istinaden değişim gösterebilir. Bu nedenle, farklı kümeleme algoritmaları farklı verilere ve farklı amaçlara uyar. Sonuç olarak kullanılması gereken en iyi kümeleme algoritması uygulamaya bağlıdır. Birçok algoritmayı deneyip, en faydalısını seçme gibi bir yaklaşım pek yaygın değildir (Joydeep Ghosh ve Alexander Liu, 2006).

K-ortalamalar algoritması, **kullanıcı tanımlı küme sayısı** (*c*) ile verilen veri kümesinin iteratif olarak kümelenmesidir. Bu algoritma pratikte çok yaygındır; uygulama ve çalıştırma açısından rahat, göreli daha hızlı ve adaptasyonu kolaydır.

Algoritma ilk olarak rastgele *c* gözlemi seçer. Bu gözlemler başlangıçta küme ortalamalarını veya merkezlerini ifade eder. Kalan her bir gözlem için gözlem ile küme ortalamaları arasındaki mesafeye göre bir kümeye atanmaktadır. Sonrasında her bir küme için yeni ortalamalar hesaplanmaktadır. Bu işlem kriter fonksiyonu yakınsayana kadar (değişmeyene kadar, yada ihmal edilebilir değişikliğe ulaşana kadar) iteratif olarak devam eder.

Minimum değişim kriteri datayı organize etmede yardımcı olan yaygın bir kriterdir. Geniş ve büyük veri kümelerinin prototipler çevresindeki verilerin dağılımını minimize eder olmalı.

R^n uzayında *N* birime sahip olduğumuzu ve *c* sayıda küme ile ilgilendiğimizi farz edelim. Gözlemler ile prototipler $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_c$ arasındaki dağılım toplamı:

$$\mathrm{Q} = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik} \, \|x_k - v_i\|^2 \qquad (4.39)$$

Burada $\| \quad \|^2 \boldsymbol{x}_i$ ve $\boldsymbol{v}_i$arasındaki öklidyen farktır. Burada önemli bileşen bölme matrisi $U = [u_{ik}]$, $i = 1,2, \dots, c$, $k = 1,2, \dots, N$ birimleri kümelere dağıtmaktadır. *U* matrisi ikili değişkenleri içermektedir.$u_{ik} = 1$ ise, k biriminin i. kümeye ait olduğunu, eğer $u_{ik} = 0$ ise aynı birimin kümeden çıkarıldığını göstermektedir.

Bölme matrisi şu şartları sağlar:

- Her bir küme tüm birimleri içermez ve boş değildir (ne tam dolu, ne tam boş):

$$0 < \sum_{k=1}^{N} \mathrm{u_{ik}} < N, i = 1, 2, \dots, c \qquad (4.2)$$

- Her birim tek bir kümeye aittir:

$$\sum_{i=1}^{C} u_{ik} = 1\,, k = 1, 2, \ldots, N \tag{4.3}$$

Bölme matrislerinin (yukarıdaki iki şartı sağlayan ikili matrisler) genel ifadesi **U** olarak tanımlanacaktır ($U \in \mathbf{U}$) . *Q*' nun minimizasyonu için, bölümlendirme matrisi ve prototipler oluşturulmalıdır.

Bu optimizasyon için birçok yaklaşım vardır.

4.2.1.1 Algoritma

Girdi: *c*: küme sayısı; *V*:*N* birimden oluşan veri seti

Çıktı: c sayıda küme yapısı

Yöntem:

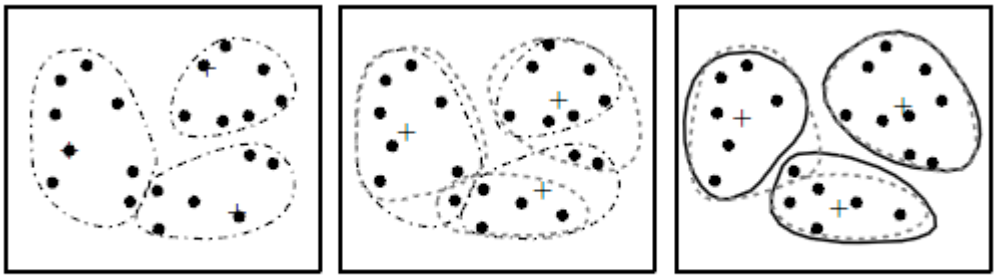
- Prototiplerin bazı başlangıç yapılanışlarıyla başlanır $v_i, i = 1, 2, \ldots, c$(rastgele seçilmiş)
- Yineleme, tekrarlama (iterate)
- Bölme matrisi oluşturma, aşağıdaki kuralla *U*' ya sayısal değer atanması. Küme ortalamalarına bakılarak benzer birimlerin kümeye yeniden atanması,

$$u_{ik} = \begin{cases} 1, & d(x_k, v_i) = min_{j \neq i} d(x_k, v_j) \\ 0, & değilse \end{cases} \tag{4.4}$$

- Ağırlıklandırılmış ortalamaları hesaplayarak, prototipler (küme ortalamaları) güncellenir, her bir küme için birimlerin ortalama değerleri hesaplanır.

$$v_i = \frac{\sum_{k=1}^{N} u_{ik} x_k}{\sum_{k=1}^{N} u_{ik}} \tag{4.5}$$

- Performans indeks Q sabitleşene ve değişmeyene kadar, yada ihmal edilebilir değişikliğe ulaşana kadar, yineleme devam eder.



Şekil 4.1 K-Ortalamalar Bölümlendirme (her bir küme ortalaması "+" ile gösterilmiştir.)

K-Ortalamalar algoritması, basitliği, etkinliği ve tarihsel öneminden dolayı birkaç yazılım programlarında (SAS FASTCLUS prosedürü ile ve Weka) yeralmaktadır.

Doğrusal complexity yapıda olması hesaplama kolaylığı sağlamasından dolayı K-ortalamaların büyük veri boyutlarında da diğer doğrusal complexity olmayan kümeleme yöntemlerine (hiyerarşik yöntemler) göre tercih edilmektedir. Complexity *i* iterasyon, *c* küme sayısı, *N* birim sayısı ve *d* değişken olmak üzere: $O = (i * c * N * d)$ olmaktadır. Normal olarak $c \ll N$ ve $i \ll N$ olacaktır.

Diğer taraftan bu algoritmanın kullanıcı tarafından tercih edilmesi yorumlaması ve uygulaması kolay olması, yakınsama hızı, dağınık ve seyrek yayılımlı verilere uyumlu olabilmesidir (Dhillon ve Modha, 2001).

K-Ortalamaların zayıf noktaları ise, başlangıç bölümlendirmesinin seçimidir. Çünkü algoritma bu seçime çok duyarlıdır, bu seçim global ve lokal minimum noktaları arasında fark oluşmasına neden olabilir. K-ortalamalar sadece izotropik küme yapılarına sahip veri kümelerinde iyi verim verebilmektedir, tek bağlantı algoritması gibi çok yönlü olamayabilmektedir. Diğer taraftan K-ortalamalar, küme ortalamaları tanımlandığı taktirde hesaplanabilmektedir. Bu da bazı uygulamalarda kategorik değişkenli verilerde gerçekleşemeyebilmektedir.

Ek olarak sapan değerlerden ve parazit (noise) verilerden etkilenebilmektedir, çünkü bu değerler hata karelerini arttırabilmektedir.

4.2.2 K-Medoid Algoritması

K-ortalamalarda hesaplanan prototipler tüm verilere dayanarak hesaplanmaktadır. Eğer herhangi bir veri parazit etkilerden oluşmaktaysa bu prototipe de etki etmektedir. Bu olumsuzlukların üstesinden gelmek için veriyi temsil eden daha sağlam yapılar oluşturulmalıdır. Bu nedenle medoid kavramı ile karşılaşmaktayız.

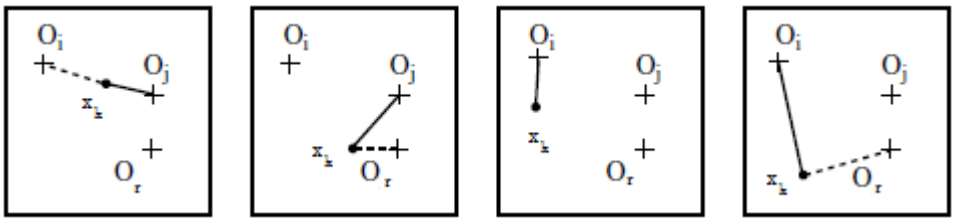
Bu nedenle hata kareler fonksiyonunu hesaplarken referans noktası olarak kümelerin ortalama değerini almak yerine her bir küme için başka bir temsili gözlemin alınması bu sorunu çözebilmektedir. Bu durumda diğer gözlemler bu küme temsili gözleme yakın, benzer olanlar şeklinde kümelenmektedir. Bu noktadan sonra bölümlendirme yöntemi, her bir noktanın referans noktalarıyla aralarındaki farklılıkların toplamının minimizasyonu ile hesaplanabilmektedir. Mutlak hata kriteri:

$$\mathrm{E} = \sum_{i=1}^{c} \sum_{k=1}^{N} |x_k - o_i| \qquad (4.6)$$

x_k gözlem değeri; o_i i. kümenin temsili gözlemi olmak üzere mutlak hatalar toplamıdır.

Başlangıç olarak temsili gözlemler (tohum) keyfi olarak seçilmektedir. İteratif işlem temsili küme gözlemleri ile temsili olmayan küme gözlemlerinin gelişmiş bir kümeleme sonucu verene kadar yer değiştirmesi ile devam etmektedir. Kaliteli kümeleme yapısı temsili küme gözlemi ile diğer gözlemler arasındaki farkın hedef fonksiyon ile minimizasyonu ile gerçekleşmektedir. Temsili olmayan o_r' yi (rastgele), mevcut o_i veya o_j temsili gözlemin iyi bir alternatifi olduğunu düşünürsek, Şekil 4.2' deki şu durumlar ile karşılaşmaktayız:

- Durum 1: x_k, o_j temsili gözlemine ait değer olsun, o_j temsili gözlemi, o_r rastlantısal gözlem ile yer değiştirirse (yeni temsili gözlem), x_k artık o_i' ye ait olacaktır $(i \neq j)$.
- Durum 2: x_k, o_j temsili gözlemine ait değer olsun, o_j temsili gözlemi, o_r rastlantısal gözlem ile yer değiştirirse (yeni temsili gözlem), x_k artık o_r' ye atanacaktır.
- Durum 3: x_k, o_i temsili gözlemine ait değer olsun, o_j temsili gözlemi, o_r rastlantısal gözlem ile yer değiştirirse (yeni temsili gözlem), x_k hala o_i temsili gözlemine ait olacaktır, bir değişim olmayacaktır.
- Durum 4: x_k, o_i temsili gözlemine ait değer olsun $(i \neq j)$, o_j temsili gözlemi, o_r rastlantısal gözlem ile yer değiştirirse (yeni temsili gözlem),x_k, o_r' ye yakın oluyor, bu nedenle bu temsili gözleme atanmaktadır.



Şekil 4.2 Temsili gözlem ile yeni gözlemin yer değiştirmesi

Yeni atamaların gerçekleştiğinde, mutlak hata E' deki fark maliyet fonksiyonuna eklenmektedir. Bu nedenle mevcut temsili gözlem, temsili olmayan gözlem ile yer değiştirdiğinde mutlak hatadaki farkı maliyet fonksiyonu hesaplamaktadır. Toplam maliyet negatif ise, o_j temsili gözlemi, o_r ile mutlak hata azalana kadar yer değiştirmektedir. Eğer toplam maliyet pozitif ise, iterasyonda değişim gerçekleşmemektedir.

4.2.3 PAM (medoid etrafında bölümlendirme)

Amaç *N* gözlemde *c* küme tanımlanmasını sağlamaktadır. İlk olarak rastgele *c* küme temsili noktaları seçilir, sonrasında algoritma daha iyi bir küme temsili noktaları geliştirir. Biri temsili gözlem olmak üzere bütün gözlem çiftleri incelenmektedir. Her bir iterasyonun karmaşık yapısı $O(c(N-c)^2)N$ ve *c*' nin büyük değerleri için hesap daha da zorlaşmaktadır.

Algoritma: Bölümlendirme temelli k-medoid algoritması

Girdi: *c*: küme sayısı; *V*: *N* birimden oluşan veri seti

Çıktı: *c* küme yapısı

Yöntem:

- *V* veri kümesinden *c* sayıda keyfi gözlem seçimi (başlangıç temsili değerler, tohum)
- Tekrarla
- Her bir gözlemin temsili gözlem sayesinde kümelere atanması
- Rastgele temsili olmayan gözlemin o_r seçilmesi
- Temsili gözlem$o_j$ileo_r' nin yer değiştirmesi sonucu hedef fonksiyonun hesaplanması
- Eğer "toplam maliyet"<0, o_j ile o_r' nin yer değiştirir.
- Değişim olmayana kadar devam eder

Ortalama ve medoid' in küme merkezleri olarak kullanılmasına karşın alternatif ölçümlerde bölümlendirmeli kümeleme yöntemlerinde kullanılmaktadır. K-ortanca ve K-mod bu alternatif ölçümler sayesinde geliştirilmiştir. Deneyimsel sonuçlar PAM' in küçük veri kümelerinde etkin olduğunu göstermiştir (Krzysztof J. Cios vd. 2007).

4.2.4 Büyük Veritabanları Bölümlendirmeli Kümeleme: CLARANS

PAM gibi K-medoid yöntemleri küçük veri kümelerinde etkili çalışmakla beraber, büyük veri kümelerinde iyi ölçümlenememektedir. Büyük veri kümelerinde örnekleme temelli, CLARA (Clustering LARge Applications) olarak adlandırılan yöntem kullanılmaktadır.

CLARA da amaç, tüm veriyi gözlemlemek yerine, tüm verileri temsil edebilecek güce sahip, verilerin daha küçük sayıda bölümünü incelemektir. CLARA çoklu örneklemin çekilmesi ile, her bir örneğe PAM uygular ve en iyi kümelemeyi çıktı olarak sunar. Dolayısı ile PAM' daki büyük verilerle çalışma konusunun üstünden gelebilir. Her bir iterasyonun complexity yapısı *s* örnek hacmi, *c* küme sayısı, *N* toplam gözlem sayısı olmak üzere$O(cs^2 + k(N - c))$ olmaktadır.

CLARA faydası örneklem hacmine bağlıdır. PAM verilen veri kümesi içinden, en iyi *k* medoid' leri seçmesine karşın, CLARA en iyi *k* medoid' leri veri kümesinden seçilen örneklem içerisinden seçmektedir. Bu nedenle eğer o_j en iyi k-medoid' lerden biriyse ve seçilen örneklemde yer almıyorsa, CLARA en iyi kümelemeyi bulamayacaktır. Eğer yanlı örnekleme söz konusu ise, örnekleme yöntemli kümeleme analizlerinde elde edilen kümeleme, tüm verilerden kümeleme kadar iyi olamayacaktır (Jiawei Han, 2006).

4.3 Hiyerarşik Yöntemler

Hiyerarşik kümeleme, ağaç kümeleme yöntemleri ile birimleri gruplamayı hedefler. Verilerin grafiksel ifadeleri ile türetilmektedir. Hiyerarşik kümeleme yöntemleri birimleri birbirleri ile değişik aşamalarda bir araya getirerek ardışık biçimde kümeler belirlemeye ve bu kümelere girecek birimleri hangi uzaklık ölçüsünde kümeye atfedileceğini belirleyen yöntemlerdir. Bu yöntemler işleyişi nedeni ile iki alt gruba ayrılmaktadır:

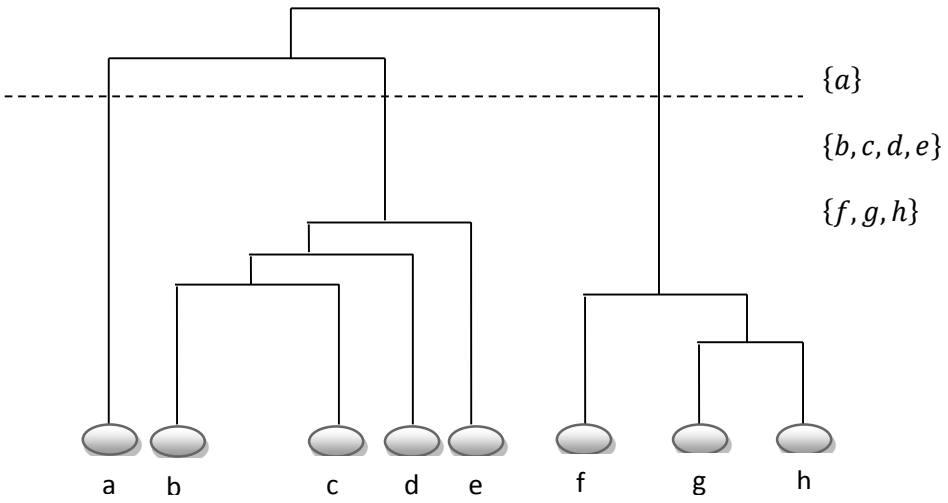
4.3.1 Birleştirici Hiyerarşik Kümeleme

Başlangıç olarak her birim kendisinin bir kümesi olarak ifade edilir. Sonrasında kümeler istenilen küme yapıları oluşturuluncaya kadar birleştirilir. Sonuçta da birimlerin tümü bir kümeye bağlanır. Bağlanma düzeyleri dendogram denilen ağaç grafiği ile gösterilir.

4.3.2 Ayırıcı Hiyerarşik Kümeleme

Birleştirici hiyerarşik yöntemin aksine, başlangıçta tüm birimleri içeren tek bir küme bulunmaktadır. Sonrasında istenilen küme yapıları oluşana kadar bölünerek alt kümeler oluşmaktadır.

Hiyerarşik yöntemlerin sonuçları dendogram ile gösterilmektedir. Dendogram gruplama değişimlerinde benzerlik veya başkalık ölçü seviyelerinde birimleri ve bulundukları grupları ağ yapısında sergilemektedir (dendogram tüm verileri tek bir köke bağlı dalları ile tutan ikili bir ağaç görünümündedir). Kümeleme işlemi ise istenilen benzerlik ölçüsünde dendogramı keserek gerçekleşmektedir.Şekil 4.3' de grafikte yukarı doğru çıkarken, bazı benzerlik ölçüleri varsayımında en yakın noktalar birleşmektedir. Örneğin burada $\{a, b, c, d, e, f, g, h\}$ veri noktaları olmak üzere, g ve h arasındaki mesafe en küçük olan mesafedir, bu nedenle bu iki birim ilk olarak birleşmiştir. Yukarı doğru çıkıldıkça kümeler büyümeye başlamaktadır. Örnekte dendogram üç kümede kesilmiştir.$\{a\}, \{b, c, d, e\}, \{f, g, h\}$



Şekil 4.3 Dendogram, verideki görsel yapı.

Hiyerarşik kümeleme yöntemleri benzerlik ölçülerinin hesaplamalarına göre de bölümlendirilmektedir:

Tek Bağlantı Kümelemesi: İki küme arasındaki mesafeyi, bir kümeden diğer kümeye aralarında en kısa mesafe olan birimlerin mesafesine eşitlemeyi göz önünde bulundurur. Aynı şekilde benzerlik söz konusu ise iki küme çifti arasındaki benzerlik, bir kümeden diğer kümeye aralarında en büyük benzerliğin bulunduğu birimler göz önünde bulundurulur (Sneath ve Sokal, 1973).

Tam Bağlantı Kümelemesi: İki küme arasında mesafenin, bir kümeden diğerine aralarında en büyük mesafe bulunan birimlerin mesafesine eşitlemeyi göz önünde bulundurmaktadır (King, 1967).

Ortalama Bağlantı Kümelemesi: iki küme arasındaki mesafeyi, bir kümeden diğerine ortalama mesafede bulunan birimlerin mesafesine eşitlemeyi göz önünde bulundurmaktadır.

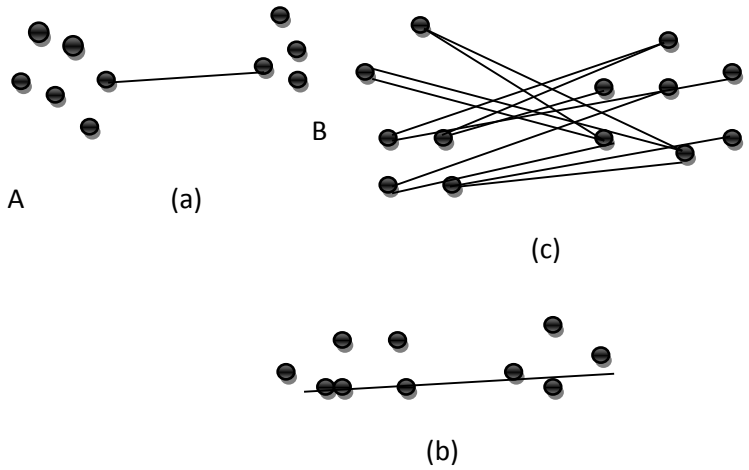
Kümeler arası mesafeler için 4 ölçü kullanılmaktadır, iki birim veya nokta arasındaki mesafe p ve p' noktalar olmak üzere $|p - p'|$' dir; m_i, $C_i$kümesinin ortalaması, n_i ise C_i deki birim sayısıdır.

Minimum mesafe : $d_{min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} |p - p'|$ (4.7)

Maksimum mesafe : $d_{max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} |p - p'|$ (4.8)

Küme Orta Nokta mesafesi : $d_{mean}(C_i, C_j) = |m_i - m_j|$ (4.9)

Ortalama Mesafe : $d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} |p - p'|$ (4.10)



Şekil 4.4 A ve B kümelerine ilişkin mesafe ölçüm yöntemleri: (a) Tek bağlantı; (b) Tam bağlantı; (c) ortalama bağlantı

Maksimum ve minimum ölçüler, kümeler arasındaki mesafe ölçülerinde iki aşırı değeri ifade etmektedir. Bu nedenle sapan veya yanlış değerlere karşı duyarlıdır. Ortalama mesafe ölçüleri minimum ve maksimum mesafelere göre sapan değerlere daha az duyarlı olmaktadır. Ortalama mesafe ölçüsü kategorik değişken tiplerinde de hesaplanabilmektedir. Fakat küme orta mesafelerinde bu durum zor veya tanımlanamamaktadır.

Hiyerarşik kümeleme yöntemleri yöntem olarak kolay olmasına karşın, birleştirme yada ayırma noktalarını seçim aşamalarında bir takım zorluklarla karşılaşmaktadır (Jiawei Han, 2001). Bu karar çok kritik bir aşamadır, çünkü bir grup birim birleştiğinde veya ayrıldığında bir sonraki adım yeni oluşmuş kümeler üzerinden gerçekleşmektedir. Bu noktadan sonra işlem **geriye alınamamakta** ve kümeler arasında birimlerin yer değiştirmesi söz konusu olamamaktadır. Bu durum doğru seçilmemiş birleştirme veya ayırma noktalarında düşük kalitede küme yapılarının oluşmasına neden olabilmektedir. Diğer taraftan yöntemin **ölçülebilmesi mümkün olamamaktadır**, çünkü her birleştirme veya ayrım kümelerin ve onları oluşturan birimlerin sayılarının doğru bir şekilde değerlendirilmesi ve incelenmesi gerektirmektedir. Diğer taraftan hiyerarşik algoritmaların complexity yapısı en azından $O = (N^2)$(burada N birim sayısı (gözlem sayısı)) yani gözlem sayısına göre doğrusal olmayan yapısından dolayı ölçümlemede yetersizlikler bulunmaktadır.

Hiyerarşik yöntemlerle kümeleme işleminin verimini arttırmak amacıyla hiyerarşik kümeleme yöntemleri ile diğer yöntemlerin entegre edilmesiyle çoklu faz kümeleme

yöntemleri geliştirilmiştir. Bunlardan BIRCH, ROCK, Chameleon yoğun kullanım alanına sahiptir.

4.4 Yoğunluk Temelli Kümeleme

Keyfi şekillere sahip küme yapıları keşfedebilmek adına yoğunluk temelli kümeleme yöntemleri geliştirilmiştir. Yoğunluk temelli kümelemeler, her bir kümeye ait noktaların belirli bir olasılık dağılımından geldiği varsayımı üzerine işleyen yöntemlerdir (Banfield ve Raftery, 1993). Tüm verilerin dağılımının birkaç dağılımın karışımından oluştuğu farzedilmektedir. Yöntemin amacı kümeleri ve onların dağılım parametrelerini belirleyebilmektir.

Süreç mevcut kümenin komşuluğunda belirli eşik değerleri aşmasında yoğunluğu kadar büyümesidir. Bir gözlemin ε yarıçapında komşuluğu, gözlemin $\varepsilon - komşuluğu$ olsun. $\varepsilon - komşuluğu$ en azından minimum sayıda gözlem içermelidir. Her bir küme yoğunluk fonksiyonunun lokal mod veya maksimumuyla nitelendirilir. Bu yöntem mod bulma olarak adlandırılmaktadır.

Bu alandaki çoğu çalışma yoğunluk bileşenin çok değişkenli Gaussian yada multinominal olduğu varsayımı temelindedir.

Bu durumda kabul edilebilir bir çözüm maksimum benzerlik yönteminin kullanılmasıdır. Bu prensibe göre, öyle kümeleme yapısı ve parametreler seçilmeli ki, böyle kümeleme yapıları ve parametrelerden türetilen verinin olasılığı maksimum olmalı. Genel olarak eksik veri problemlerinde uygulanana "beklenen maksimizasyon algoritması" (**EM**) (Dempster vd., 1977) parametre tahmin problemlerinde de kullanılmaktadır. Algoritma parametre vektörünün başlangıç vektörünün başlangıç tahmini ile başlamakta ve sonrasında iki aşamayı takip eder (Farley ve Raftery, 1998): adım-E (**beklenen**), koşullu beklenti, benzerlik ve mevcut parametre tahminleri hesaplanır; adım-M (**maksimizasyon**), parametreler, adım-E' de tanımlanan beklenen benzerliği maksimize etmektedir. Bu algoritma gözlemlenen verinin benzerliğinin lokal maksimuma yakınsadığını göstermektedir.

DBSCAN(parazit verilerle, yoğunluk temelli uzamsal kümeleme yöntemi) algoritması keyfi şekilli kümelerin keşfedilmesi ve büyük uzamsal veri kümelerinde etkili bir yoğunluk temelli kümeleme yöntemidir. Algoritma veri tabanlarında her bir gözlemin

komşuluğunu araştırır ve minimum sayıda gözlemi içerip içermediğini kontrol eder(Ester vd. 1996).

OTOSINIF (AutoClass) sık kullanılan EM algoritmasının da bir türevi olan bayesgil kümeleme yöntemidir ve birçok değişik dağılımı kapsayan (Gaussian, Bernoulli, Poisson, and log-normal dağılımlar) (Cheeseman ve Stutz, 1996) algoritmadır.

4.5 Izgara Temelli Kümeleme (GRID)

Bu algoritma gözlem uzayını, kümeleme işlemi için tüm faaliyetlerin gerçekleştiği oluşan ızgara yapılı bir şekil üzerinde sınırlı sayıda hücrelere bölümlendirmektedir. En büyük avantajı hızlı olmasıdır, çünkü genellikle gözlem sayısından bağımsızdır. STING, WaveCluster, CLIQUE belli başlı algoritmalardır.

4.6 Model Temelli Kümeleme

Model temelli kümeleme yöntemleri gözlem verileri ile bazı matematiksel modelleri uydurmayı amaçlamaktadır. Temelinde olasılık dağılımlarının karışımıyla oluşan verilerin varsayımına dayanmaktadır. Sınıf ağaçları ve sinir ağları yoğun kullanılan yöntemlerdir.

BÖLÜM 5

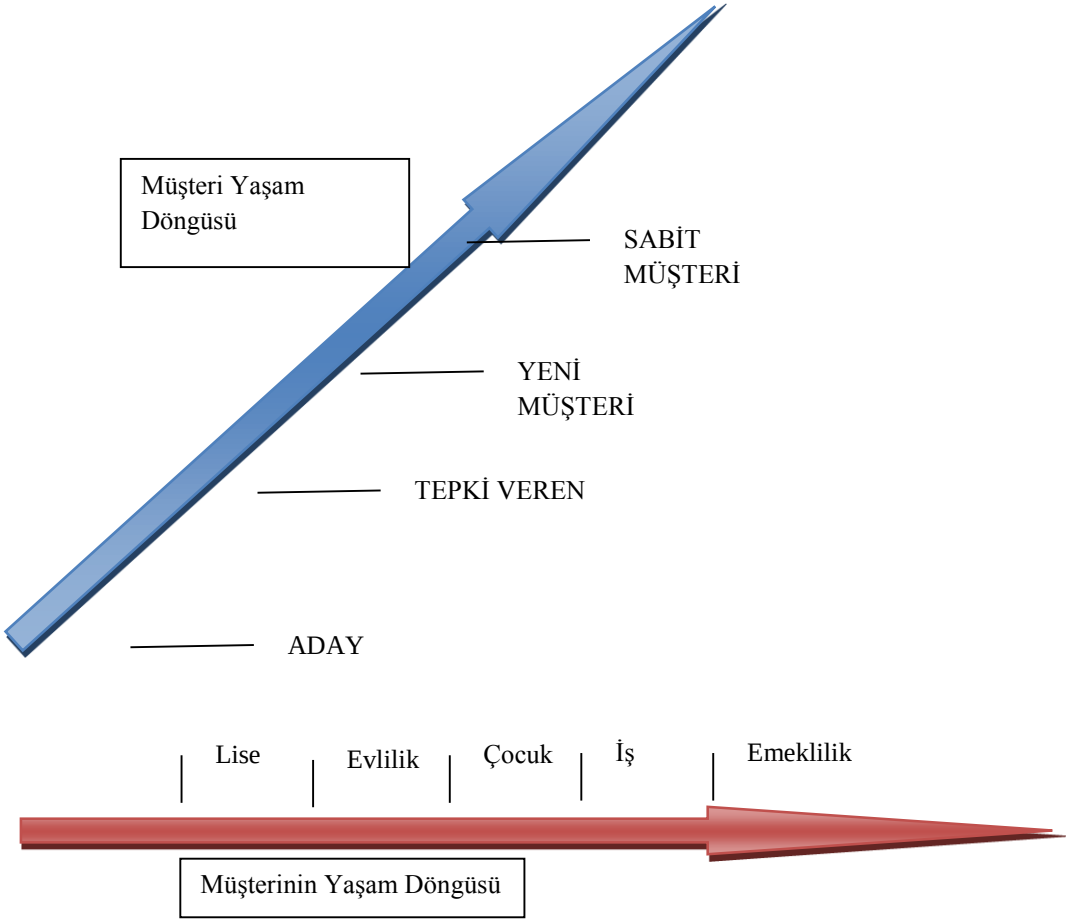
MÜŞTERİ İLİŞKİLERİ YÖNETİMİ ve PAZARLAMADA VERİ MADENCİLİĞİ YÖNTEMLERİ

5.1 Yaşam Döngüleri

Neredeyse tüm işletmeler için müşteri en önemli değer durumundadır. Müşteri firma ilişkileri zaman içerisinde çok sıklıkla değişim gösterebilmektedir. Her bir sektörde bu ilişkileri yönetebilmek için farklı yaklaşımlar mevcuttur. Bazıları için servis kalitesi, bazıları için fiyat, bazıları için pratiklik bu ilişkiyi sağlamaktadır. Müşteri etkileşimi için birçok kanal kullanılabilmektedir; doğrudan pazarlama, arama merkezleri, yüz yüze ve kitle reklam gibi. Bunun yanında web de artık müşteri ile etkileşimi sağlamada önemli bir kanal olmuştur, yakın zamanda da mobil teknolojiler de yeni bir etkileşim kanalı olarak önemli bir konuma sahip olacaktır.

Teknolojinin işletmelerin müşterileri ile olan etkileşimlerinde, birebir insan insana etkileşime oranla daha kişisel ve pazarlama odaklı yeralabileceği söylenebilir.

Bu nedenle müşterinin işletme ile ilişkisinin başlangıç ve bitişi süreci daha anlamlı olmaktadır. Müşteri yaşam döngüsü işletme açısından müşteri evreleri ve müşterinin kendi sosyo-ekonomik yaşamındaki evreler ile birlikte değerlendirilmelidir. Bu açıdan "Müşteri Yaşam Döngüsü" işletme açısından müşteri yaşam döngüsünü, "Müşterinin Yaşam Döngüsü" müşterinin sosyo-ekonomik durumundaki değişimleri ifade edecektir.



Şekil 5.1 Müşteri yaşam döngüleri ve müşteri yaşam evreleri

5.1.1 Müşteri Yaşam Döngüsü

Perakendeciler ve müşteri ile ilişkili firmalar için "müşteri yaşam döngüsü" uzun dönem "müşteri değeri" sağlayabilme açısından çok önemlidir.

Müşteri değerini arttırabilmek için;

- Sahip olunan müşterilerin satın alımlarının (hizmet veya ürün) arttırılması
- Daha yüksek marjlı ürün satmak
- Yada onları uzun periyotta elde tutabilmek

şeklinde üç başlıkta toplanabilir.

Değerli denebilecek müşteri sürekli aynı değere sahip olmamalıdır. İlişki mutlaka zaman içerisinde değişim gösterir. Bu ilişkiyi anlayabilmek Müşteri İlişkileri Yönetimi (CRM)' in en önemli bölümüdür. Müşteri yaşam döngüsü içerisindeki evreleri tespit (tahmin) edebilmek Veri Madenciliği' nin en çok başvurulduğu konudur.

Müşteri deneyimi terim olarak firmalara çok anlam ifade etmelidir. Müşteri deneyimi firmayla müşteri arasındaki ilişkinin pratikte müşteri tarafındaki algısıdır. Kurum ile müşteri etkileşiminin detayını ifade etmektedir.

Müşteri yaşam döngüsünün doğru algılanması, veri madenciliğinin daha etkili sonuçlar sağlamasına yardımcı olmaktadır. Yaşam evrelerinin ve olaylarının tahmini bir şirketin vazgeçemeyeceği değerde önemli bir bilgidir.

Müşteriyi anlamaya yönelik diğer bir boyut ise müşteri **yaşam döngüsüdür**. Müşteri yaşam döngüsü, müşteri ile işletme arasındaki ilişkinin zaman içindeki değişimini inceler. Her sektör ve işletme için farklı yaşam döngüleri tanımlansa da, beş ana fazdan oluşmaktadır (Şekil 5.1);

Adaylar, hedef pazarda yer alan henüz müşteri olmayan potansiyel kitledir. Hedef Pazar firmaların satmış oldukları ürün veya verdikleri hizmete göre farklılık teşkil edebilmektedir. Bir sigorta şirketi için Hedef Pazar sağlıklı insanların veya düşük risk grubundaki insanların hepsi olabileceği gibi, bir lüks giyim firması için Hedef Pazar AB grubu sosyo-ekonomik durumda yer alan insanlar olabilecektir.

Firmaların ürün veya hizmetlerine ilgi duyabileceğini düşündükleri adaylara (Hedef Pazar) yönelik yapılan pazarlama kampanyaları, yeni müşteri kazanım kampanyalarıdır. Henüz müşteri ile ilgili hiçbir veri olmamasına karşın, benzer kampanyalardan faydalanmış veya benzer iletişimlere tepki vermiş müşteri profilleri oluşturulabilir, hatta bugünün karlı müşterilerinin profiline yakın olabilecek adaylar Pazar Araştırmaları ile saptanabilir ve reklam veya diğer medya mecraları duyuru için kullanılmaktadır.

Tepki verenler, programa cevap veren üyeler, işletme ile bir aktivite, web üzerinden veya mobil aygıtlar ile kayıt veya başvuru gerçekleştirmiş adaylardır. Hedef Pazarda bulunan adaylar için gerçekleştirilen yeni müşteri kazanım kampanyalarına cevap veren müşterilerdir. Henüz aktivasyon gerçekleştirmemiş olmalarına karşın programa üye olan müşterilerdir.

Adayların “Tepki Veren” olabilmesi

- Web üzerinden kayıt alınarak
- Direk firma üzerinden uygulamaya kayıt olarak
- Telefon ile kayıt alınarak

gerçekleşmektedir.

Hangi adayların tepki veren müşteri olacağı ve hangi tepki veren müşterinin yeni ve sabit müşteri olacağına dair tahmin modelleri yine VM çatısında irdelenmektedir.

Yeni müşteriler, programa cevap veren üyelerin işletmeden alışveriş yapmaları, bir anlaşmayı imzalamaları veya kişisel bilgilerini paylaşımı gerçekleştiren müşterilerdir. Bu müşteriler artık firmanın üretmiş olduğu ürün veya hizmeti kullanmakta olan müşteridir. Müşteri ilk satın alımını gerçekleştirmiştir ve firma için yeni müşteridir.

Sabit müşteri, zamanla tekrar gelen bu yeni müşterilerin işletme ile daha geniş ve derin ilişkiler oluşturması ile gerçekleşen fazdır. Firma ile müşteri arasında ekonomik ve manevi ilişkilerin gerçekleşmesiyle artık tepki veren veya yeni müşterilerin sabit müşteri olduğu söylenebilir. İlk satın alımını gerçekleştiren müşteri artık firma için potansiyel sabit müşteridir. Bu evreden sonra gerçekleşen olaylar;

- Müşterilerin satın alım veya kullanım alışkanlıkları müşteri davranışlarını tanımlayabilmektedir. Farklı segment gruplarında farklı kullanım alışkanlıkları ortaya çıkabilmektedir.
- Çapraz satış kampanyaları müşterilere ürün veya hizmet satın alımında daha geniş bir ilişki sağlayabilir.
- Yine üst satış ile var olan ürün veya hizmetin bir üst seviyede sunulması müşteriyle olan ilişkiyi daha da derinleştirebilir.

Sabit müşteri evresi firmalar için en önemli evredir. Müşterinin hangi ürünü veya hizmeti satın aldığı ve ürün veya hizmetle ilgili tüketim süreci Müşteri Davranış yapısının algılanabilmesi için önemli fikir sağlayabilmektedir. Hangi müşterilerin çapraz satışa, hangi müşterilerin üst satışa tepki verebileceğini tahmin edebilmek firmalar için oldukça değerlidir.

Müşteri davranışı her işlemin detaylı kayıt edildiği geniş veritabanlarından ortaya çıkmaktadır.

Kayıp müşteriler, bir şekilde işletme ile bağını koparan müşteri kitlesidir. Müşterinin işletme ile bağını koparmasının birkaç sebebi olabilir; isteğe bağlı kayıp (müşteri rakip işletmeye yönelebilir veya artık müşteri için ürün değerini yitirmiş olabilir), zorunlu kayıp (Kredi kartı borcunu ödemeyen bir müşteri artık o banka için zorunlu olarak eski müşteri

olacaktır.), beklenen kayıp (müşteri artık hedef pazar içerisinde yeralmayabilir, müşterinin taşınması gibi…)

Temelde iki farklı kayıp bulunmaktadır. İlki isteğe bağlı (gönüllü) kayıp, sabit olan müşteri artık müşteri olmak istemiyor. Neden?

Müşteri, firmanın veya firmanın bayisinin olmadığı bir yere taşınmış olabilir

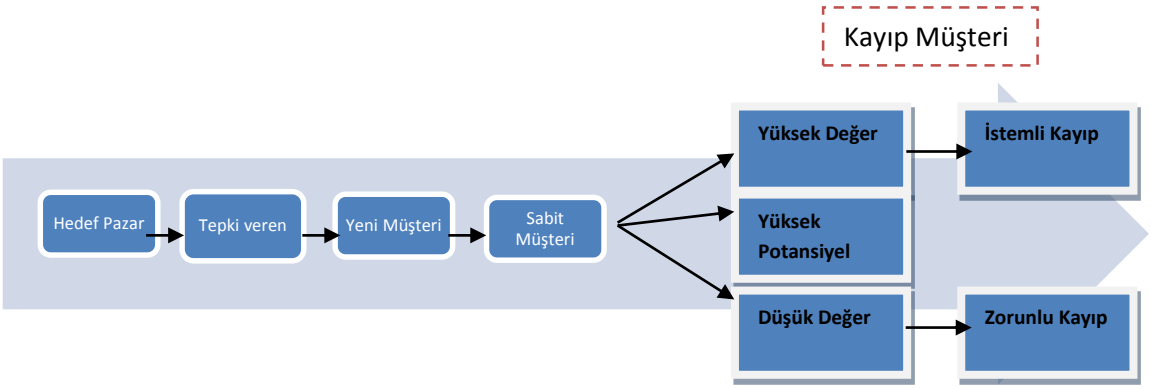
- Yaşam döngüsü gereği artık firmanın ürünlerine ihtiyacı olmayabilir
- Pazarda rakip firmanın avantajı sebebiyle kayıp olabilir
- Müşteri artık ürünü veya hizmeti kullanmanın bir değerini görmüyor olabilir

Sabit müşteri ne zaman kayıp olacak? Bu müşterinin yaşam-zaman değeri (YZD) ile alakalıdır. Bir müşteri ne kadar uzun süreli müşteri ise o kadar değerli müşteridir.

Gönülsüz kayıp, bunlar bazı sebeplerden ötürü satın alım alışkanlıklarını düşüren, eskiye göre daha az gelen, daha az harcayan müşterilerdir.

Bu müşteriler henüz kayıp sayılmaz eski alışveriş alışkanlıklarına kazandırmak için geri kazanım kampanyaları gerçekleştirilebilir.

Çok katlı bir giyim mağazası için aday, kıyafet ihtiyacı olan ve satmakta olduğu ürün çeşidi profiline uyan herhangi biri olabilir; programa cevap veren, gelip mağazayı ziyaret eden; yeni müşteri, ilk defa alışveriş yapan veya sadakat programına üye olan; sabit müşteri ise tekrar alışveriş gerçekleştiren; eski müşteri ise, uzun bir süredir mağazaya gelmeyen müşteridir. Her bir sektör için tanımlar değişim gösterebilir.



Şekil 5.2 Farklı evreler içerisindeki müşteri yaşam döngü süreci

5.1.2 Müşteri Yaşam Döngüsü İşletme Süreçleri

Müşteri yaşam döngüsü müşterileri işletmeyle ola ilişkilerinin derinliği ve süresi açısından tanımlayabilmektedir. Bu döngüde yeralan işletme süreçleri ise müşterileri bir yaşam döngüsü fazından diğerine taşımaktadır (Şekil 5.2). Bu süreçleri takip etmek işletme için çok değerlidir, çünkü işletmeler müşterilerini zaman içerisinde daha değerli yapmayı amaçlamaktadırlar. Bu süreçlerin ortaya çıkarılmasında VM rolü de çok değerlidir.

5.1.3 Müşteri Kazanım

Müşteri kazanımı, hedef pazarda bulunan adayları bir şekilde etkileyerek müşteriye dönüştürmektir. Bu çoğunlukla reklam, sosyal medya veya kişiden kişiye bilgi aktarımı gibi pazarlama ve reklam iletişimleri ile gerçekleşmektedir. Müşteri kazanımı sırasında üç soru karşımıza çıkmaktadır: Adaylar kim? Bir müşteri ne zaman kazanıldı? VM' nin kazanımdaki rolü ne?

Adayların kim olduğunu bilmek önemli, çünkü iletişim mesajları bu adaylara hedeflenmektedir.

- Coğrafi genişleme işletmeye aday kazandırır, bu adaylar mevcut müşteriye benzerlikler gösterebildiği gibi, göstermeyebilir
- Ürün, servis ve fiyatta gerçekleşen değişiklikler farklı segmentler getirebilir
- Rekabet gereği hedef kitlenin değiştirilmesi zorunluluğuna gidilebilir

Müşterinin ne zaman kazanıldığına dair çıkarımlar;

- Müşteri bir şekilde bir zamanda cevap verir. Bu satış tarihidir.
- Hesap&Kayıt temelli ilişkilerde müşteri hesap oluşturur. Bu hesap açılış tarihidir.
- Bir şekilde hesabın kullanıldığı tarih

Mevcut veriler tahmin modellerini kısıtlayabilmektedir. Tahmin modelleri iletişim maliyetlerinin göreli yüksek olduğu doğrudan posta ve tele-pazarlama gibi kanallar için kullanılmaktadır. Amaç cevap vermesi muhtemel ve iyi bir müşteri olacağı tahmin edilen adaylarla temasları sınırlayabilmek.

- Aday kaynağı
- Bireysel ve hane halkı verilerinin eklenmesi
- Coğrafi anlamda demografik verilerin eklenmesi

5.1.4 Aktivasyon

Adayların bir şekilde ilgi göstermeleriyle bir takım aktivasyon süreçleri başlamaktadır. Aktivasyon bir kayıt formunu doldurma ile gerçekleşebileceği gibi, daha ötesinde bir onaylama süreci de olabilir. Aktivasyon analitik ihtiyaçlardan ziyade işletme ihtiyaçlarına odaklanan operasyonel süreçtir. Müşteri aktivasyonu müşteri ilişkisinin öncül koşullarını sağlamaktadır. Bu öncül koşullar uzun dönem müşteri davranışlarının faydalı göstergeleridir.

5.1.5 İlişki Yönetimi

Aday artık müşteri olduktan sonra amaç müşteri değerini arttırabilmektir. Bu da bazı uygulamalar ile gerçekleşebilmektedir;

Üst Satış, müşterilerin daha zenginleştirilmiş içerikli ürünleri almalarını sağlamak

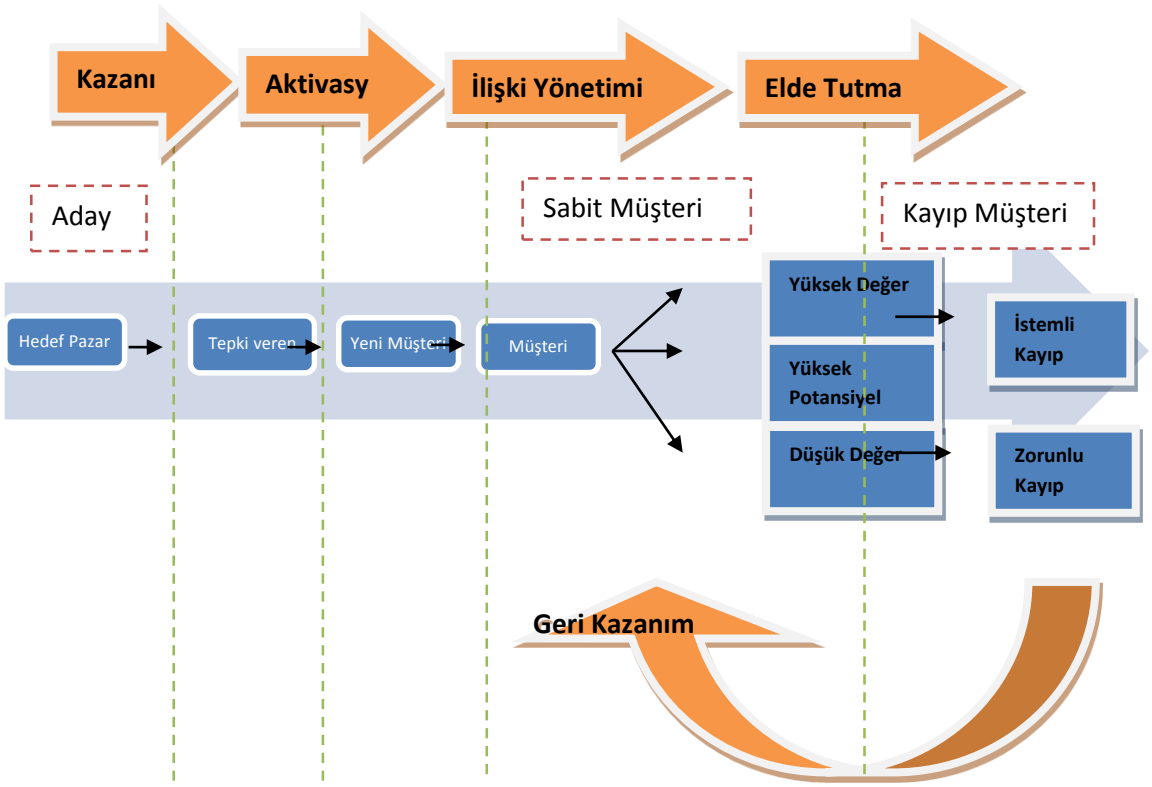
Çapraz Satış, müşteri ilişkilerinin genişletilmesi, CD ve uçak bileti satın alan müşterilere ek olarak kitap önermek

5.1.6 Elde Tutma

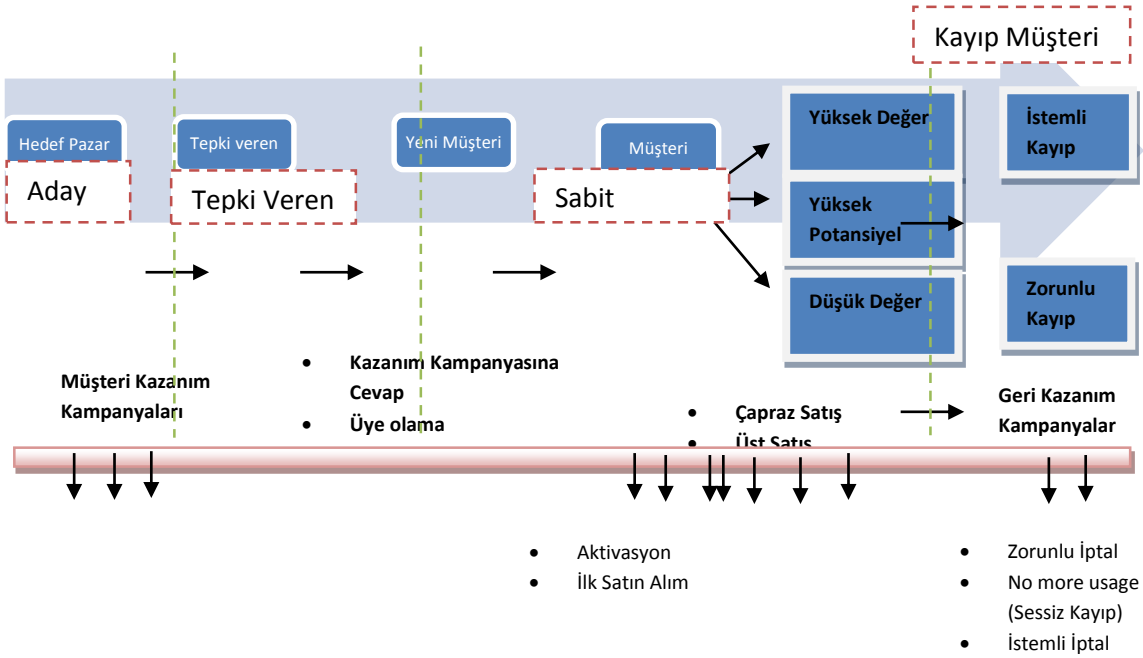
Müşteri elde tutma tahmin modellerinin uygulandığı alanlardan biridir. Sağkalım analizi müşterilerin müşteri olma sürelerini anlamaya yönelik parametreler ortaya çıkarmaktadır. Sağkalım analizi belirli bir zaman sonra müşterinin kaybedileceğine dair olasılıklar atayabilmektedir.

5.1.7 Geri Kazanım

Müşteri gitmiş olsa bile hala geri gelme olasılığı bulunmaktadır. Geri kazanım değerli olan müşterileri geri getirebilmek için bir takım kampanya, promosyon, fiyat indirimleri, yeni ürünler ve hediye ürün önerileriyle sağlanabilmektedir. Geri kazanım veri analizinden daha çok operasyonel staratejilere bağlıdır. Geri kazanım aslında elde tutma çalışmaları ile yakından ilişkilidir. Müşterilerin neden firmayla ilişkilerinin koptuğu bilgisi müşteriyi elde tutmaya yönelik gelecek çalışmalara yön verecektir.



Şekil 5.3 Müşteri yaşam döngüsü içerisinde yer alan işletme süreçleri



Şekil 5.4 İşletme aksiyonları

5.1.8 Müşterinin Yaşam Döngüsü: Yaşam Evreleri

Firma tarafından müşterini yaşam döngüsünü gördük, ama kendimizi müşteri yerine koyarsak dinamikler biraz daha farklı olmaktadır. Müşterinin kendisinin de bir yaşam döngüsü bulunmaktadır. Öyle ki onlarında hayatlarında müşteri olarak değerlerine etki eden olaylar mevcuttur. Bu olaylar, mevcut olan müşterinin ne zaman kayıp olacağını veya hangi ürüne ihtiyaçları olduğunu belirleyebilir. Nedir bu olaylar?

- İş değişikliği
- İlk çocuk
- Evlilik
- Boşanma
- Emeklilik
- Taşınma
- Büyük hastalıklar

Bu ve benzeri bilgiler oldukça kişisel fakat firmalar için bir o kadarda önemli bilgilerdir. Bu bilgileri güncel tutabilmek firmalar için müşteri yaşam döngüsünü anlayabilme açısından oldukça değerlidir.

Müşterinin yaşam döngüsü, her bir bireysel müşterinin yaşamlarındaki dönüm noktalarını oluşturan olayları içermektedir. Müşteri yaşam döngüsü içerisinde farklı evreler mevcuttur. Bu evreler işletme ile müşteriler arasındaki ekonomik ilişkinin anlaşılmasına yönelik durumun ortaya çıkmasını sağlar ve müşteri ilişkilerini tanımlar. Müşteri yaşam döngüsü müşteri davranışlarının anlaşılabilmesi için bir iskelet oluşturmaktadır.

Bu yaşam döngüsü unsurları birçok olayı içerebilir üniversitede okumak, yeni bir işe başlamak, evlenmek, çocuk sahibi olmak, emekli olmak gibi evreler olabileceği gibi; ilk arabaya sahip olmak, piyango biletinin vurması gibi olayları da içerebilen kişiyi veya kişinin ekonomik süreci etkileyen unsurlardır. Örneğin “taşınmak” insan hayatındaki anlamlı bir olaydır. Bir insan taşındığında yeni mobilyalar satın alır, yeni bir banka hesabı açar, yerel gazeteye abone olur. Dolayısı ile kimin taşındığını bilmek mobilya satıcısı, yerel gazete ve banka için bu bireyleri hedefleyebilmek açısından yararlıdır. Bu ve benzeri yaşam evrelerini anlamak işletmelerin ürün tanımlarını yapmaları ve hangi müşteri ile ne tür iletişim dilinin kullanacağının belirlenmesinde yardımcı olmaktadır.

Çoğu büyük işletmeler müşterilerin yaşam evreleri bilgilerini bilmek ve buna göre ürün geliştirmek ve pazarlama mesajlarını zenginleştirmek istemektedirler, fakat müşterinin yaşamında gerçekleşen bu olayları bilmek çok kolay olmamakta ve işletmeler çoğunlukla müşterilerin kendileri ile paylaştıkları bu bilgilerin değerini bilememekteler. Örnek olarak, bir müşteri taşındığında ve bu bilgiyi işletme ile paylaştığında müşterinin yeni adresi eski adresi ile yer değiştirmekte, oysa her iki adresi tutmak işletmeye önemli bir avantaj sunabilir. Benzer şekilde bayan müşteriler evlendiklerinde soy isimlerinde yaptıkları değişiklikleri işletme ile paylaşabilmekteler. Bu bilgi o müşterinin artık çift olduğunu ve finansal durumunun eskisinden farklı olacağının bilgisidir. Dolayısıyla bu bilgiyi kayıt altına almayan işletmeler kendileri için kazançlı olacak hedef mesajı müşterileri için oluşturamamaktadırlar.

Yaşam evreleri açısından müşteri ilişkilerinin yönetimindeki zorluklar,

- Yaşam evrelerini oluşturan olayları zamanında belirleyebilmek,
- Olayların bir çoğunun bir defa veya çok nadir gerçekleşmesi,
- Yaşam evreleri olayları tahmin edilemez ve işletme kontrolü dışında gerçekleşir.

Çizelge 5.1 Yaşam döngüsü evreleri ve davranış alışkanlıkları

Yaşam Döngüsünde Evreler	**Satın Alım ve Davranış Alışkanlıkları**
1. Eğitim& Öğrenim Evresi: genç, yalnız insan evde yaşamıyor	Kısıtlı düşük finansal gelir. Moda takipçisi. Eğlenceye yönelik, genellikle klasik mutfak harcaması, temel eşyalar, ev eşyası, araba, hobi malzemeleri veya gezi ihtiyaçları bulunmaktadır
2. Yeni evli çiftler: genç ve henüz çocuk sahibi değil	Yakın geleceğe göre finansal olarak daha iyi durumda. Yüksek harcama oranı ve ağırlıklı dayanıklı tüketim malları, genellikle araba, buzdolabı, dayanıklı ev eşyaları, tatil ihtiyaçları bulunmaktadır
3. Dolu Yuva I: 6 yaş altı çocuk sahibi	Ev harcaması en yüksek, nakit varlık az. Biriken para ve finansal konumdan dolayı tatminsiz. Yeni ürünlere ilgili, genellikle TV, bebek ürünleri, yiyecek, soğuk algınlığı ilaçları, vitaminler, oyuncak, kızak, kayak, bisiklet ihtiyaçları bulunmaktadır
4. Dolu Yuva II: 6 ve üzeri yaş çocuk sahibi	Finansal olarak daha iyi durumda. Aile işleri. Reklamdan az etkilenen, harcama sepeti büyük, genellikle bir çok yiyecek, temizlik maddesi ve bisiklet satın alır. Müzik dersleri, piyano…

Çizelge 5.1 Yaşam döngüsü evreleri ve davranış alışkanlıkları…(devamı)

Yaşam Döngüsünde Evreler	**Satın Alım ve Davranış Alışkanlıkları**
5. Dolu Yuva III: daha yaşlı çiftler birden fazla çocuk sahibi	Finansal olarak hala iyi. Daha çok aile işi. Bazı çocuklar iş sahibi. Reklamlarla etkilenmesi çok zor. Yüksek dayanıklı tüketim malları harcaması. Genellikle yeni, daha hoş eşyalar, araba seyahatleri, gereksiz aletler, aksesuarlar, kayık tekne sevdası, diş hizmetleri, dergiler.
6. Boş Yuva I: daha yaşlı çiftler, çocuklarıyla yaşamayan	Ev sahipliği üst seviyede. Biriken gelir ve finansal durumdan memnun. İlgi alanları gezmek, dinlenmek, kişisel eğitim. Yardım ve bağış yapar. Yeni ürünlere karşı ilgisizdir. Genellikle, tatil ve lüks ev düzenlemeleri yaparlar
7. Boş Yuva II: çocukları ile beraber yaşamayan emekli çiftler	Gelirde şiddetli düşüş. Evde yaşama. Genellikle medikal cihazlar, sağlığa uyumaya ve sindirime iyi gelecek medikal ürünler.
8. Yalnız yaşayan, çalışan	Gelir hala iyi, fakat muhtemelen evini satmayı düşünmekte
9. Yalnız yaşayan, emekli	Diğer emekli grup gibi bazı medikal ve ürün ihtiyaçları olmaktadır. Gelirde büyük düşüş. Nezaket, bağlılık, ve güvenlik anlamında özel ihtiyaçlar.

BÖLÜM 6

PAZARLAMADA SAĞKALIM ANALİZİ

Sağkalım analiz, aynı zamanda "olay zamanı" analizi olarak da adlandırılmaktadır. Müşteriyi anlama adına oldukça değerli olan sağkalım analizinin adından da anlaşılacağı gibi terminolojisinin ve köklerinin medikal araştırmalardan ve üretimdeki hata analizlerinden gelmesine karşın pazarlamada da kullanılmaktadır. Sağkalım pazarlamada müşterinin ne zaman ilişkiyi sonlandıracağından endişelenmek gerektiği bilgisini verebilmektedir. Her ne kadar müşteri sözel olarak işletmeyle konuşmasa da veritabanına kayıtlı veriler işletmelere bir şeyler söylemek istediklerini açıklayan yapılar sunabilmektedir. Amaç müşteri kaybını ve alışveriş alışkanlıklarının düşmesini engellemek ve nedenlerini araştırmaktır.

Müşteri sadakati bütün işletmeler için oldukça değerlidir. İşletmeler değerli müşterileri ile uzun süreli ilişki kurmak ve müşterilerin işletmeye kattıkları değeri arttırmak isterler. Bu doğrultuda bu müşterilerin ileriki herhangi bir zamanda işletme ile ilişkilerini koparmalarının tahmin edilmesi işletmeye ileride gerçekleşecek değer kaybını gösterebilmektedir ve buna yönelik aksiyonların alınmasını sağlayabilmektedir.

Sağkalım analizinin önemi olay zamanı karakteristiklerini anlamayı sağlar,

- Müşteri ne zaman ilişkiyi koparacak
- Bir sonraki zaman müşteri farklı bir segmente göç edebileceği
- Bir sonraki zaman müşteri işletmeyle olan ilişkileri daha da derinleştirecek veya daha da daraltabileceği
- Müşteri ilişkilerindeki faktörler muhtemel müşteri olma sürelerini (tenure) arttırabilir veya azaltabilir.

- Müşteri olma sürelerinde çeşitli faktörlerin sayısal etkisi

  Bu bilgiler pazarlama süreçlerini besleyen bilgilerdir.

6.1 Müşteriyi Elde Tutma

Müşteriyi elde tutmaya yönelik gerçekleştirilen çalışmalar aslında sağkalım analizinin bir başlangıcı ve kaba bir yaklaşımı olabilir. Müşteriyi elde tutmayı anlayabilmek için müşteriye dair iki bilgiyi bilmek gerekir:

- Ne zaman müşteri oldukları
- Ne zaman müşteri olmayı bıraktıkları

Bu iki değer arasındaki fark, müşteri olma süresini verecektir ve müşteriyi elde tutmaya yönelik iyi bir ölçüdür. Bu ölçümü yapabilmek işletmeler için göründüğü gibi kolay olamayabilir, ilk önce başlangıç ve bitişlere karar vermek gerekir, bu karar işletmenin tipine ve mevcut veriye bağlıdır.

Örneğin bir gazete veya dergi üyeliğinde başlangıç ve bitiş tarihleri nettir (abonelik temelli ilişki; sigorta, iletişim, kablolu televizyon, gazete/dergi üyeliği, banka, elektrik). Müşteri belirli bir zamanda üye olur ve ödemeye başlar, ne zaman ödemeyi keser ve üyeliği devam etmezse süreç sonlanır. Banka hesapları, sigorta poliçeleri, elektrik, kablolu tv servis üyeliği belirli bir zamanda başlar ve başka bir zamanda da sona erer. Bu nedenle bu tür işletmelerde müşteri olma süreleri daha iyi tanımlanabilmektedir.

Diğer taraftan bazı işletmelerin bu şekilde sürekli ilişkileri olmayabilir. İşlem bazlı çalışan işletmelerde özellikle perakende, web portalları gibi, müşteri alışverişleri (ziyaretleri) zamana yayılmış veya bir kez ise durum biraz daha farklıdır. Başlangıç zamanı belli, genellikle ilk alışveriş veya ilk ziyarettir. Son ise daha karışıktır, bir takım işletme kuralları atanarak elde edilebilmektedir. Örnek olarak 12 aydır işletmeyi ziyaret etmeyen veya alışveriş gerçekleştirmemiş müşteri artık işletme ile ilişkisini bırakmış müşteri olarak tanımlanabilir. Müşteriyi elde tutma analizleri bu tanımlamalarda oldukça faydalı bilgiler sunabilmektedir.

Hesaplanan müşteri olma süreleri, eğri üzerinde işaretlenebilmektedir. Bu eğri belirli bir zamanda elde tutulan müşterilerin oranını göstermektedir.

6.2 Hazard

Elde tutma eğrileri veriyi anlamada ne kadar yararlıda olsa genel bir şekli, parametrik bir formu ve teorisi bulunmamaktadır.

Risk olasılıkları bu düşünceyi biraz daha açabilmektedir. Amprik hazard olasılığı, önceden belirlenmiş bir formu veriye uydurmadan geçmiş verilere bakıp neler olup bittiğini anlamayı sağlar. Ayrıca müşteriyi elde tutmaya yönelik bilgiler sağlar ve elde tutma eğrilerini düzenleyerek sağkalım eğrileri oluşturmaktadır.

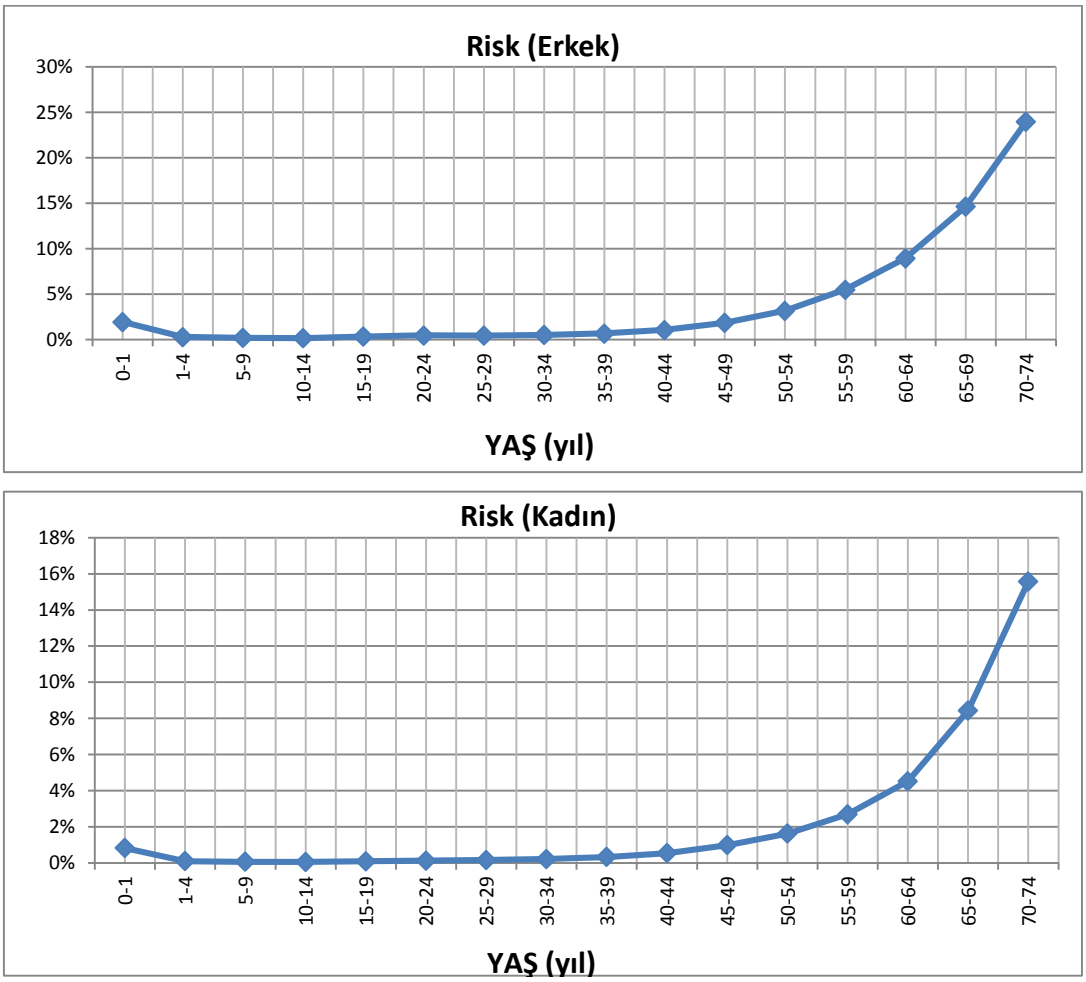
Belirli bir zaman aralığında sahip olunan müşterinin müşteri olma süresi t olsun. Müşterinin $t + 1$ zamanda müşterinin ayrılma olasılığı nedir? Sorusuna cevabı hazard olasılığı verebilmektedir. Hazard t ve $t + 1$ zaman arasında müşteriyi kaybetme riskidir. Tanımlamanın daha net olabilmesi için birinin belirli bir yaş aralığında ölüm oranları, hayat tablolarını gözlemlersek;

(Şekil 6.1) Türkiye Hayat ve Hayat Annüite Tablolarının Oluşturulması Projesi Sonuç Raporuna, 1927-2000 nüfus sayımlarına göre Türkiye Hayat Tablolarının erkek ve kadın ayrı olarak grafiksel gösterebilmektedir. Hayat tabloları risk için iyi bir örnek sunmaktadır. Grafiğin karakteristik şeklinden dolayı banyo küveti şekli de denilmektedir. Risk yüksek başlamakta ve belirli uzun bir periyod aynı kalıp, daha sonra giderek artmaktadır.

Her ne kadar müşteri riskleri (risk) yıl yerine gün, hafta veya ay olarak hesaplansa da, aynı yaklaşım müşteri olma süreleri içinde kullanılabilir. Bu durumda t müşteri olma süresinde risk hesabı iki bilgiye gereksinim duyar. İlki t zamanda (ya da t ve $t + 1$ zaman arasında) duran müşteri sayısı. İkincisi ise bu zaman içerisinde durabilecek müşteri sayısı, risk altındaki kitle de denilebilir. Dolayısı ile bu müşteri olma zamanı t ve daha fazla olan müşterileri, aynı zamanda t zamanda ilişkisini koparan müşterileri de içermektedir. Risk olasılığı bu iki değerin oranıdır ve olasılık olmasından dolayı 0 ile 1 arasında değer almaktadır. SAS ve SPSS gibi istatistiksel yazılımlarda bu risk hesaplamaları hayat tabloları fonksiyonları ile gerçekleşebilmektedir. Fakat direk müşteri veritabanında da bu hesaplamalar gerçekleşebilmektedir.

Hesaplamanın doğru olabilmesi için, popülasyonda yer alan herhangi bir bireyin belirli bir zamanda durma şansının olmalıdır. Bu hesaplama yönteminden ziyade, riski hesaplamada kullanılacak verini özelliğidir.

Eğer popülasyon yeterince büyükse, veriler istatistiksel olarak güvenirlilik ve standart hata gibi kavramlarda sıkıntı oluşturmayacaktır.



Şekil 6.1 Türkiye Hayat ve Hayat Annüite Tablolarının Oluşturulması Projesi Sonuç Raporuna, 1927-2000 nüfus sayımlarına göre Türkiye Hayat Tabloları

6.2.1 Risk(Hazard) Fonksiyonu Örnekleri

6.2.1.1 Sabit Risk Fonksiyonu

Müşterilerin ne kadar süredir işletmeyle ilişkileri olduğu irdelemeden, müşterilerin ayrılma riskinin aynı olduğu varsayılır. Örneğin günlük ölçülen riskin %0,1 olduğunu düşünelim. Bu her bin müşteriden birinin bir günde işletmeyi bırakacağını söyler.

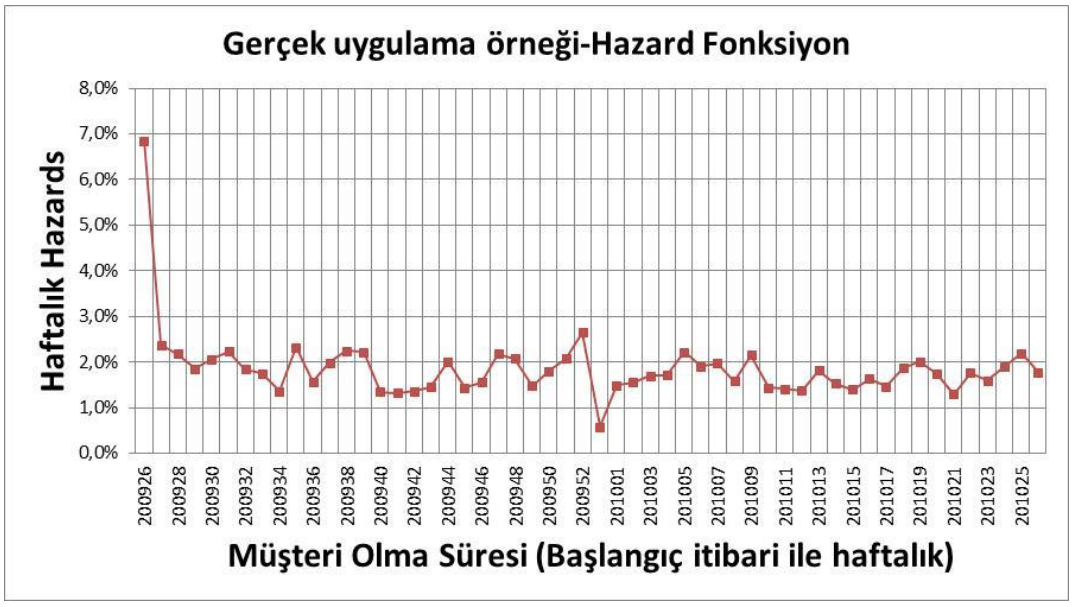
6.2.1.2 Banyo Küveti Şekilli Risk Fonksiyonu

Hayat tabloları banyo küveti şeklindeki risk fonksiyonlarına bir örnektir. Risk ilk olarak biraz yüksek başlar daha sonra düşer ve uzun bir süre düzenli devam eder, sonunda da artar. Anlaşmalı üyelik veya satın alımlarda (anlaşmalı cep telefonu gsm üyeliği, adsl üyeliği), anlaşmanın hemen sonrasında ciddi bir düşüş yaşanır, çünkü müşteri ya servisin iyi olmadığını düşünür yada ödeme gerçekleştirmez. Sonrasında anlaşma süresince müşteri ya firma tarafından caydırılır ya finansal cezalardan dolayı yada anlaşmanın şartlarından dolayı kullanımı zorunlu hisseder. Anlaşma bitiminden sonra müşteriler çoğunlukla ayrılma eğilimlidir ve oranlar giderek yükselir, çünkü artık müşteri anlaşması bittikten sonra daha rahat olur. Bu noktada ürün veya hizmet fiyatı rakiplerine göre avantajlı olmayabilir. Pazar değişebilir ve müşteri bu değişikliğe cevap verir.

6.2.1.3 Gerçek Yaşamdan Örnek

Şekil 6.2' de gerçek bir uygulama örneği yer almaktadır. Çok katlı mağazacılık alanında perakendenin önde gelen şirketlerinden birinin 2009 yılı ilk 6 ayında kazanılan yeni müşterilerin (programa üye olan) haftalık risk fonksiyonu incelenmiştir. Her ne kadar abonelik temelli bir müşteri ilişkisi olmasa da, çünkü başlangıç ve bitiş tarihleri perakende de abonelik sistemlerindeki kadar açık olmayabiliyor. Risk fonksiyonu haftalar sonra müşterilerin ilişkilerini kesme olasılıklarını gösterebilecektir.

Risk eğrisinin ilginç karakteristik özellikleri bulunmaktadır. İlk olarak yüksek başlar, bunun nedeni yeni müşteri elde edinimi için gerçekleşen hedef kampanyanın süresinin bitmesi, önemli bir işbirliğinin bitmesi veya sürekli bir sadakat programının sona erdirilmesi gibi nedenlerdir. Daha sonrasında promosyon, indirim odaklı kitlesel kampanyalar ve özel günlerde 2009 35. hafta(okul öncesi dönem) ve 52-53. haftalar (yılbaşı dönemi); 2010 19. hafta (Anneler Günü) ve 25. hafta (Babalar Günü) pik değerler görülebilmekte fakat giderek azalan bir eğim göstermektedir.

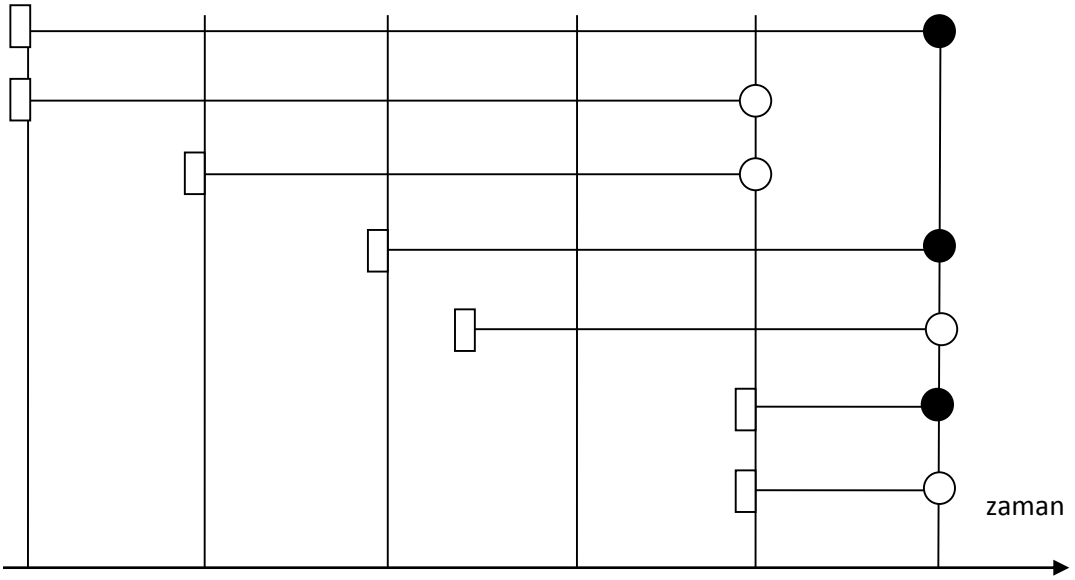


Şekil 6.2 Gerçek değerlerle risk fonksiyonu

6.3 Sansürleme

Risk olasılığı, göre t zamanda durma sayısının, popülasyona bölünmesi ile elde edilmektedir. Bu esnada t zamanından önce duran müşteriler, popülasyon içerisinde olmamaları gerekmektedir. Bu durum bireysel olarak t zamanda sağkalım olasılığını koşullu yapmaktadır.

Sansürlemeye başka bir örnek ise, müşteri olma süreleri t olan fakat aktif olan bir müşteriler risk için t müşteri olma süresinde popülasyona dahil edilmezler, çünkü $t + 1$' de gidebilirler. Bu müşteriler t' den küçük değerlerde risk hesaplamasında yer alabilirken t' de risk hesaplamalarına girmezler.



Şekil 6.3 Farklı zamanlarda başlayan müşteri grupları, bazı müşteriler hala aktif oldukları için sansürlenmiştir.

Şekil 6.3' de farklı tarihlerde ilişkileri başlayan müşteriler için bitiş açık veya kapalı küçük halkalarla gösterilmektedir. Eğer halka açıksa, müşterinin bırakmış ilişkiyi olduğunu anlıyoruz ve bitiş tarihinin de belli olmasıyla müşteri olma süresini biliyoruz demektir.

Eğer halka kapalı ise, bitiş tarihi henüz belli değil, bu müşteri veya müşterinin müşteri olma süreleri sansürlendi anlamına gelmektedir. Müşterinin kesin ayrılış tarihinin bilinmemesinden dolayı minimum müşteri olma süresi ancak söylenebilir. Şekil 6.3' den müşteri olma sürelerinin tablosu aşağıdaki gibidir:

Çizelge 6.1 Müşteri olma süreleri

1	evet	5
2	hayır	4
3	hayır	3
4	evet	3
5	hayır	2
6	evet	1
7	hayır	1

Çizelge 6.2 Zaman periyodunda müşteri takibi

Müşteri	Sansürlenme	Yaşam Süresi	T0	T1	T2	T3	T4	T5
1	evet	5	aktif	aktif	aktif	aktif	aktif	aktif
2	hayır	4	aktif	aktif	aktif	aktif	pasif	sansür
3	hayır	3	aktif	aktif	aktif	pasif	sansür	sansür
4	evet	3	aktif	aktif	aktif	aktif	sansür	sansür
5	hayır	2	aktif	aktif	pasif	sansür	sansür	sansür
6	evet	1	aktif	aktif	sansür	sansür	sansür	sansür
7	hayır	1	aktif	pasif	sansür	sansür	sansür	sansür

Çizelge 6.3 Risk olasılıkları

	T0	T1	T2	T3	T4	T5
Aktif	7	6	4	3	1	1
Pasif	0	1	1	1	1	0
Sansür	0	0	2	3	5	5
Risk	%0	%14	%20	%25	%50	%0

Diğer bir sansürleme örneği de rekabetçi sansürlemedir. İki türlü kayıp (pasif) müşteriden bahsedebiliriz: isteğe bağlı ayrılık ve isteğe bağlı olmadan ayrılık. İsteğe bağlı ayrılık, müşterinin ayrılma kararından kaynaklanır, isteğe bağlı olmayan ise kurumun kararından kaynaklanır, genelde yapılmayan hesap ödemelerinden kaynaklanır.

Hazard değerleri müşterinin belirli bir zamanda müşteri olmayabileceğine dair bir olasılık sunmaktadır. Sağkalım ise müşterinin bu zamana kadar müşteri olabileceği olasılığını bildirmektedir. Bu nedenle sağkalım değeri tam olarak hazard değerlerinden hesaplanmaktadır.

Zamanın bir noktasında, müşterinin bir sonraki zamanda sağkalma şansı *1- Hazard* olacaktır, bu t zamanda şartlı sağkalım olacaktır. Şartlı sağkalım olmasının nedeni, t zamana kadar sağkaldığı farzedildiğinden kaynaklanmaktadır.

Sağkalım eğrileri müşteri elde tutma eğrilerine göre müşteri elde tutma amaçlarında daha erken sonuçlar verebilmektedir. Müşteri elde tutma eğrilerinde zamana göre daha sıçramalı bir yapı oluşurken, sağkalım eğrilerinde daha düzgün bir yapı seyrelmektedir.

- Sağkalım hesaplaması tüm verileri kullandığı için, elde tutma hesaplarına göre daha istikrarlıdır.
- Sağkalım eğrileri aşağı doğru eğimli oldukları için, müşteri yarı ömür değerleri ve müşterinin ortalama yaşam sürelerinin hesaplanması daha doğru olacaktır.
- Daha fazla bilgi dahil edilmesiyle sağkalım daha doğru ve düzgün yapıda müşteri elde tutma örüntüleri sağlayacaktır.

Müşterileri analiz ederken risk ve sağkalım müşteriler hakkında anlamlı ve değerli bilgiler vermektedir. Sağkalımın kümülatif olması, farklı grup müşterileri karşılaştırmak için iyi bir özet bilgi sunabilmektedir.

Diğer taraftan sağkalım eğrileri kümülatif olduğu için, zamanın belirli bir anındaki yapıları görmek zor olacaktır. Fakat hazard ile bu nedenler daha açıkça görülebilmektedir. Diğer taraftan farklı gruplar için hazard kıyaslaması yapmak doğru olmayacaktır. Çünkü matematiksel olarak hesaplanmış ortalama hazard anlamlı olmayacaktır.

Buraya kadar zamanın kesikli değişken olarak düşünülmektedir. Gün içerisinde özellikle bir zaman dikkate alınmamıştır. Fakat zamanın sürekli olduğu varsayımında da bu hesaplamaların gerçekleşebilmesi gerekmektedir. Bu durumda hazard olasılığı, yerine hazard oranları hesaplanmaktadır ve sonrasında sağkalım eğrileri oluşmaktadır.

Kesikli zaman verilerinde çok iyi işleyen bu hazard hesaplamalarına aynı zamanda hayat tablosu yöntemi de denmektedir. Benzer olarak sürekli zaman verilerinde de Kaplan-Meier yöntemi uygulanmaktadır. Her iki yöntemin kesikli zaman verilerinde uygulanması benzer sonuçları vermektedir.

İstatistiksel sağkalım analizinin önemli bir bölümü ise parametrelenmiş regresyon ile hazard tahmini yapmaktır.

Sağkalım analizi için, birçok yöntem bulunmaktadır;

Hayat tabloları,

Kaplan-Meier tahmincileri,

Exponential regresyon,

log-normal regresyon,

Oransal hazards regresyon,

Rekabetçi Risk Modelleri,

Ve Kesik zamanlı yöntemler bunlardan birkaçıdır.

Sürekli rassal değişken *T,* kümülatif dağılım fonksiyonu $F(t)$ ve olasılık yoğunluk fonksiyonu (oyf)$f(t)$ olsun. $F(t)$ aynı zamanda hata fonksiyonu olarak da bilinmektedir.

Hata fonksiyonu (kdf);

$$\Pr(T \leq t) = F(t) \quad (6.1)$$

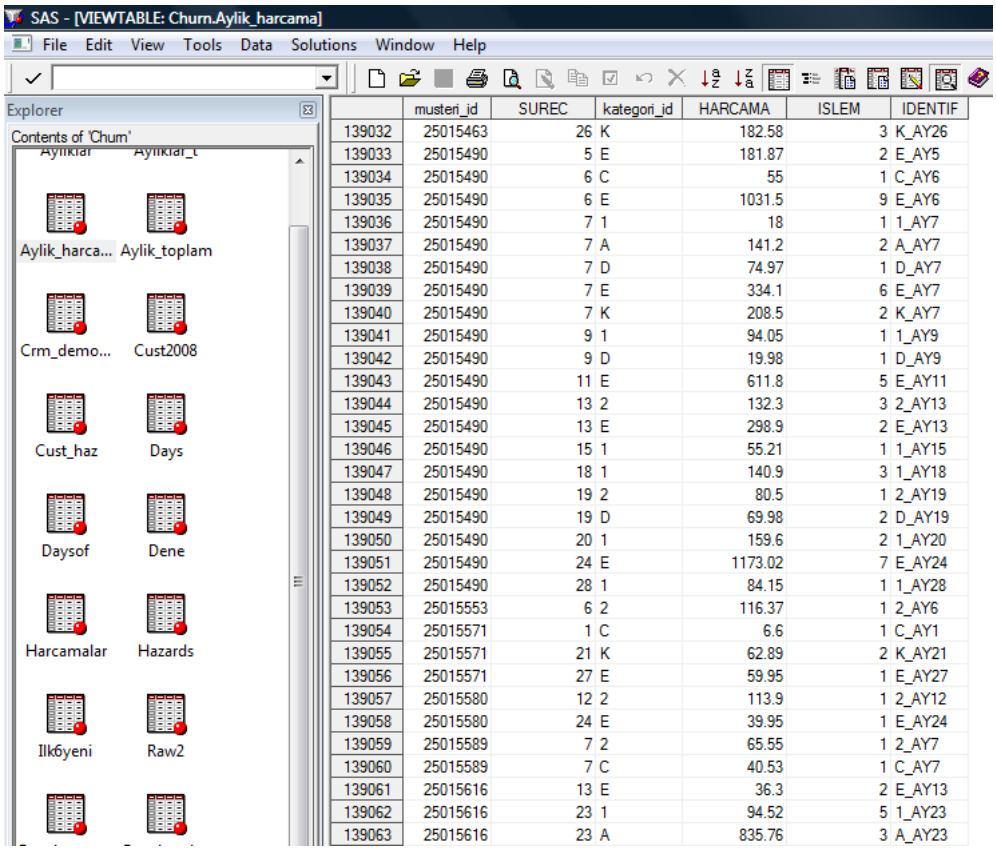
Sağkalım fonksiyonu;

$$\Pr(T > t) = 1 - F(t) \equiv S(t) \quad (6.2)$$

Sağkalım Analizi perakende içinde churn modellemelerinde kullanılmaktadır. Müşterisinin geçmiş eğilimlerinden de faydalanarak, ne kadar daha müşteri olarak kalacağı, alışkanlığın değişip değişmediğini öngörülebilmektedir.

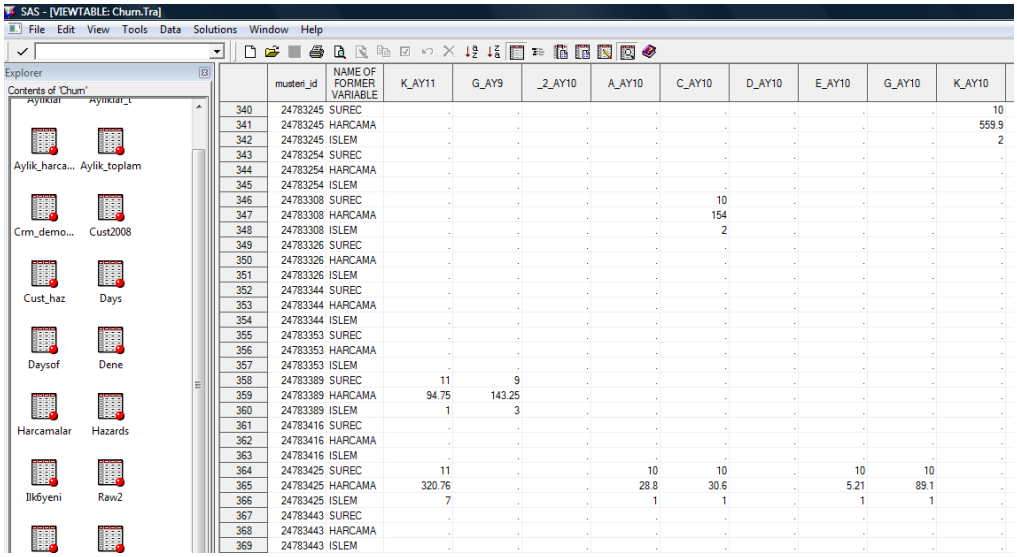
Çizelge 6.4' deki işlem verileri üzerinden müşteri geliş alışkanlıkları ve harcama trendleri oluşturulabilmektedir.

Çizelge 6.4 Müşteri Aylık harcama tutarları/kategorik



	musteri_id	SUREC	kategori_id	HARCAMA	ISLEM	IDENTIF
139032	25015463	26	K	182.58	3	K_AY26
139033	25015490	5	E	181.87	2	E_AY5
139034	25015490	6	C	55	1	C_AY6
139035	25015490	6	E	1031.5	9	E_AY6
139036	25015490	7	1	18	1	1_AY7
139037	25015490	7	A	141.2	2	A_AY7
139038	25015490	7	D	74.97	1	D_AY7
139039	25015490	7	E	334.1	6	E_AY7
139040	25015490	7	K	208.5	2	K_AY7
139041	25015490	9	1	94.05	1	1_AY9
139042	25015490	9	D	19.98	1	D_AY9
139043	25015490	11	E	611.8	5	E_AY11
139044	25015490	13	2	132.3	3	2_AY13
139045	25015490	13	E	298.9	2	E_AY13
139046	25015490	15	1	55.21	1	1_AY15
139047	25015490	18	1	140.9	3	1_AY18
139048	25015490	19	2	80.5	1	2_AY19
139049	25015490	19	D	69.98	2	D_AY19
139050	25015490	20	1	159.6	2	1_AY20
139051	25015490	24	E	1173.02	7	E_AY24
139052	25015490	28	1	84.15	1	1_AY28
139053	25015553	6	2	116.37	1	2_AY6
139054	25015571	1	C	6.6	1	C_AY1
139055	25015571	21	K	62.89	2	K_AY21
139056	25015571	27	E	59.95	1	E_AY27
139057	25015580	12	2	113.9	1	2_AY12
139058	25015580	24	E	39.95	1	E_AY24
139059	25015589	7	2	65.55	1	2_AY7
139060	25015589	7	C	40.53	1	C_AY7
139061	25015616	13	E	36.3	2	E_AY13
139062	25015616	23	1	94.52	5	1_AY23
139063	25015616	23	A	835.76	3	A_AY23

Çizelge 6.5 Aylık Harcama ve İşlem bilgileri/kategorik



	musteri_id	NAME OF FORMER VARIABLE	K_AY11	G_AY9	_2_AY10	A_AY10	C_AY10	D_AY10	E_AY10	G_AY10	K_AY10
340	24783245	SUREC									10
341	24783245	HARCAMA									559.9
342	24783245	ISLEM									2
343	24783254	SUREC									
344	24783254	HARCAMA									
345	24783254	ISLEM									
346	24783308	SUREC					10				
347	24783308	HARCAMA					154				
348	24783308	ISLEM					2				
349	24783326	SUREC									
350	24783326	HARCAMA									
351	24783326	ISLEM									
352	24783344	SUREC									
353	24783344	HARCAMA									
354	24783344	ISLEM									
355	24783353	SUREC									
356	24783353	HARCAMA									
357	24783353	ISLEM									
358	24783389	SUREC	11	9							
359	24783389	HARCAMA	94.75	143.25							
360	24783389	ISLEM	1	3							
361	24783416	SUREC									
362	24783416	HARCAMA									
363	24783416	ISLEM									
364	24783425	SUREC	11			10	10		10	10	
365	24783425	HARCAMA	320.76			28.8	30.6		5.21	89.1	
366	24783425	ISLEM	7			1	1		1	1	
367	24783443	SUREC									
368	24783443	HARCAMA									
369	24783443	ISLEM									

1. Kategori kırılımlı ilk alışveriş ve son alışveriş tarihlerin oluşturulması;

DATA CHURN.TIMES2;SETCHURN.TIMES;

ILKT_ ERKEK =INTCK('MONTH',&FIRSTDATE,ILKT_ERKEK)+**1**;

ILKT_ AYAKKABI =INTCK('MONTH',&FIRSTDATE,ILKT_AYAKKABI)+**1**;

ILKT_ COCUK =INTCK('MONTH',&FIRSTDATE,ILKT_COCUK)+**1**;

ILKT_ GAS =INTCK('MONTH',&FIRSTDATE,ILKT_GAS)+**1**;

ILKT_ KADIN =INTCK('MONTH',&FIRSTDATE,ILKT_KADIN)+**1**;

ILKT_ EV =INTCK('MONTH',&FIRSTDATE,ILKT_EV)+**1**;

ILKT_ KOZMETIK =INTCK('MONTH',&FIRSTDATE,ILKT_KOZMETIK)+**1**;

SONT_ ERKEK =INTCK('MONTH',&FIRSTDATE,SONT_ ERKEK)+**1**;

SONT_ AYAKKABI =INTCK('MONTH',&FIRSTDATE,SONT_ AYAKKABI)+**1**;

SONT_ COCUK =INTCK('MONTH',&FIRSTDATE,SONT_ COCUK)+**1**;

SONT_GAS=INTCK('MONTH',&FIRSTDATE,SONT_GAS)+**1**;

SONT_ KADIN =INTCK('MONTH',&FIRSTDATE,SONT_KADIN)+**1**;

SONT_ EV =INTCK('MONTH',&FIRSTDATE,SONT_EV)+**1**;

SONT_ KOZMETIK =INTCK('MONTH',&FIRSTDATE,SONT_KOZMETIK)+**1**;

SON=INTCK('MONTH',&FIRSTDATE,SONTARIH)+**1**;

ILK=INTCK('MONTH',&FIRSTDATE,ILKTARIH)+**1**;

FORMAT ILK **4.**;FORMAT SON **4.**;FORMAT ILKT_ERKEK **4.**;FORMAT ILKT_COCUK **4.**;

FORMAT ILKT_KADIN **4.**;FORMAT ILKT_GAS **4.**;FORMAT ILKT_AYAKKABI **4.**;FORMAT ILKT_EV **4.**;

FORMAT ILKT_KOZMETIK **4.**;FORMAT SONT_ERKEK **4.**;FORMAT SONT_COCUK **4.**;FORMAT SONT_KADIN **4.**;

FORMAT SONT_GAS **4.**;FORMAT SONT_AYAKKABI **4.**;FORMAT SONT_EV **4.**;FORMAT SONT_KOZMETIK **4.**;

KEEP MID ILK SON ILKT_ERKEK ILKT_COCUK ILKT_KADIN ILKT_EV ILKT_KOZMETIK ILKT_GAS ILKT_AYAKKABI

SONT_ERKEK SONT_COCUK SONT_KADIN SONT_EV SONT_KOZMETIK SONT_GAS SONT_AYAKKABI;**RUN**;

2. Herbir müşterinin ortalama geliş periyodlarının hesaplanması

DATACHURN.FREKANS(KEEP=MID TARIH INDIRIMLITUTAR);

SET CHURN.RAW2;

RUN;

PROCMEANSDATA=CHURN.FREKANSNOPRINT;

OUTPUTOUT=CHURN.FREKANS(DROP=_TYPE_ _FREQ_) N(TARIH)=FARKLI_TAR;

CLASS MID TARIH;WAYS**2**;**RUN**;

DATACHURN.FREKANS;SET CHURN.FREKANS;

DROP FARKLI_TAR;**RUN**;

PROCTRANSPOSEDATA=CHURN.FREKANSOUT=CHURN.FREK4;

BYMID;VAR TARIH;**RUN**;

```
DATA CHURN.FREK5;SET CHURN.FREK4;
ARRAY COL(730) COL1-COL730;
ARRAY FAR(729) FAR1-FAR729;
DO I=1TO729;
J=I+1;
IF COL(J) NE .THEN FAR(I)=INTCK('DAY',COL(I),COL(J));
ELSEIF COL(J)=.THEN FAR(I)=.;
END;RUN;

DATA CHURN.FREK5;SET CHURN.FREK5;
DROP COL1--COL730 I J;
ARRAY FAR(729) FAR1-FAR729;
DO I=1TO729;
FAR(I)=FAR(I)/30;END;RUN;
DATA CHURN.FREK7;SET CHURN.FREK5;
ARRAY FAR(729) FAR1-FAR729;
ORT=MEAN(OF FAR1--FAR729);
MINI=MIN(OF FAR1--FAR729);
MAXI=MAX(OF FAR1--FAR729);
KEEP MID ORT MINI MAXI;
RUN;
```

3. Müşterinin gelmediği sürenin hesaplanması

```
DATACHURN.DATES;SET CHURN.TIMES;
ALL_ARA_G=INTCK('DAY',SON,&LASTDATE);
```

```
A_ARA_G=INTCK('DAY',SONT_A,&LASTDATE);

C_ARA_G=INTCK('DAY',SONT_C,&LASTDATE);

E_ARA_G=INTCK('DAY',SONT_E,&LASTDATE);

G_ARA_G=INTCK('DAY',SONT_G,&LASTDATE);

K_ARA_G=INTCK('DAY',SONT_K,&LASTDATE);

_1_ARA_G=INTCK('DAY',SONT_1,&LASTDATE);

_2_ARA_G=INTCK('DAY',SONT_2,&LASTDATE);RUN;


DATACHURN.DATES;SET CHURN.DATES;

ALL_ARA_G=ALL_ARA_G/30;

_1_ARA_G=_1_ARA_G/30;

_2_ARA_G=_2_ARA_G/30;

A_ARA_G=A_ARA_G/30;

C_ARA_G=C_ARA_G/30;

E_ARA_G=E_ARA_G/30;

G_ARA_G=G_ARA_G/30;

K_ARA_G=K_ARA_G/30;

KEEP MID ALL_ARA_G _1_ARA_G _2_ARA_G A_ARA_G C_ARA_G E_ARA_G G_ARA_G
K_ARA_G ;RUN;
```

4. 1 tölerans limitli kategorilendirme

```
DATACHURN.TANIMLAR;SET CHURN.TANIMLAR;

DEG_ALL2=MEAN_ALL+STD_ALL;

DEG_12=MEAN_1+STD_1;

DEG_22=MEAN_2+STD_2;

DEG_A2=MEAN_A+STD_A;

DEG_C2=MEAN_C+STD_C;
```

DEG_E2=MEAN_E+STD_E;

DEG_G2=MEAN_G+STD_G;

DEG_K2=MEAN_K+STD_K;

IF DOLU_ALL=**.** THEN PAS_ALL2=**9**;

ELSEIF DOLU_ALL NE **.** AND DEG_ALL2 LE ALL_ARA_G THEN PAS_ALL2=**1**;

ELSEIF DOLU_ALL NE **.** AND DEG_ALL2 GT ALL_ARA_G THEN PAS_ALL2=**0**;

IF DOLU_1=**.** THEN PAS_12=**9**;

Bu şekilde müşterilerimizi genel eğilim gereği 1 tölerans limiti gösterdikleri sapma ile "pasif müşteri" olarak atamış olduk.

Bu uygulama basit ve genel anlamda herhangi bir müşterinin genel eğilimde olup olmadığını, genel trend' e göre sapan müşterilerin pasif veya aktifliğini belirler.

Çizelge 6.6 Pasiflik Sınırları

KATEGORİ	EV	KOZMETİK	AYAKKABI	ÇOCUK	ERKEK	GAS	KADIN	GENEL
Pasifleşme Sınırları	9,1	8,6	9,7	8,2	8,5	8,6	8,2	6,4

Müşterilerin kategori kırılımsız, genel olarak ortalama olarak 6,4 aydır gelmemeleri ile pasif müşteri oldukları belirlenmiştir.

Diğer taraftan, ev kategorsinde pasifleşme sınırı 9,1 ay olarak belirlenmiştir.

Bu şekilde, müşterilerin bugüne kadar gelmediği zamanlar için bir sınır belirlenmiştir. Bu sınırı aşan müşterilere churn aksiyonlarının alınması gerekmektedir.

Çizelge 6.7 Pasif Müşteri Bilgileri / Segment

Segmentler	A-çoklu	B-çoklu	C-kozmetik	D-ev	E-erkek	F-gas	G-ayakkabı	H-kadın	Total
Kişi Sayısı	6.707	30.271	29.536	15.469	28.965	37.073	16.075	26.736	190.899
%	3,51%	15,86%	15,47%	8,10%	15,17%	19,42%	8,42%	14,01%	100%
Segment Oran	12,87%	27,62%	34,54%	35,67%	32,60%	31,05%	34,19%	31,06%	26,83%

Çizelge 6.8 Aktif Müşteri Bilgileri / Segment

Segmentler	A-çoklu	B-çoklu	C-kozmetik	D-ev	E-erkek	F-gas	G-ayakkabı	H-kadın	Total
Kişi Sayısı	45.189	63.302	30.232	15.349	25.497	43.148	14.797	31.179	293.949
%	15,37%	21,54%	10,28%	5,22%	8,67%	14,68%	5,03%	10,61%	100,00%
Segment Oran	86,70%	57,76%	35,35%	35,39%	28,69%	36,14%	31,48%	36,22%	41,31%

- İlk olarak her müşterinin kendine has davranışı olabilir, dolayısı ile anlamlı bir çıktı sunmayabilir
- Hesaplanabilmesi için minimum 2 periyod olması gerekiyor

Diğer taraftan bir olasılık yaklaşımı da sağlayabiliriz;

DATA CHURN.STOC1;SET CHURN.STOC1;

PROB_1=EXP((-**1**/ORT_1)*_1_ARA_G);

KEEP MID PROB_1;**RUN**;

DATACHURN.LIMITLI_G;SET CHURN.METOD1;

IF F_GUN_G LT **3**THEN TAHMIN_G=**9**;

ELSEIF F_GUN_G GE **3**THENDO;

IF PROB_G LE **0.1** AND SON_AY_G=**0**THEN TAHMIN_G=**1**;

ELSE TAHMIN_G=**0**;END;

Eksponansiyel model ile müşteri geliş tahminine bir olasılık belirlenebilmektedir. Bu modelde, olasılık değeri %10 hata payına kadar müşteriyi kaybedebileceğimiz bir modeldir.

BÖLÜM 7

PERAKENDE' DE VERİ MADENCİLİĞİNDE KULLANILACAK VERİ YAPISI

Veri olmadan veri madenciliği yapılamayacağı için, müşteri karakteristiğini ortaya çıkarabilecek verileri (girdi) veri madenciliği algoritmalarının temel ihtiyacıdır. İşletme ihtiyaçlarının belirlenmesinden sonra bu amaçlara hitap edecek verilere ve bu verilerin derlenmesine ihtiyaç bulunmaktadır.

Perakende sektöründe firmalar müşterilerine ilişkin geniş ölçüde veriye sahiptirler. Bu veriler, satış noktalarından toplanan **işlem verileri**, müşterilerden **elde edilen bilgiler** olabilmektedir. Müşteri tanımlamaları için veri madenciliği süreçlerinin çalışacağı, müşteri segmentasyon modeli geliştirmede kullanılacak beş önemli veri tipi bulunmaktadır.

- Ürün verisi
- İşlemsel verisi
- Demografik veriler
- Müşteri ilişkileri verileri
- Ek veriler

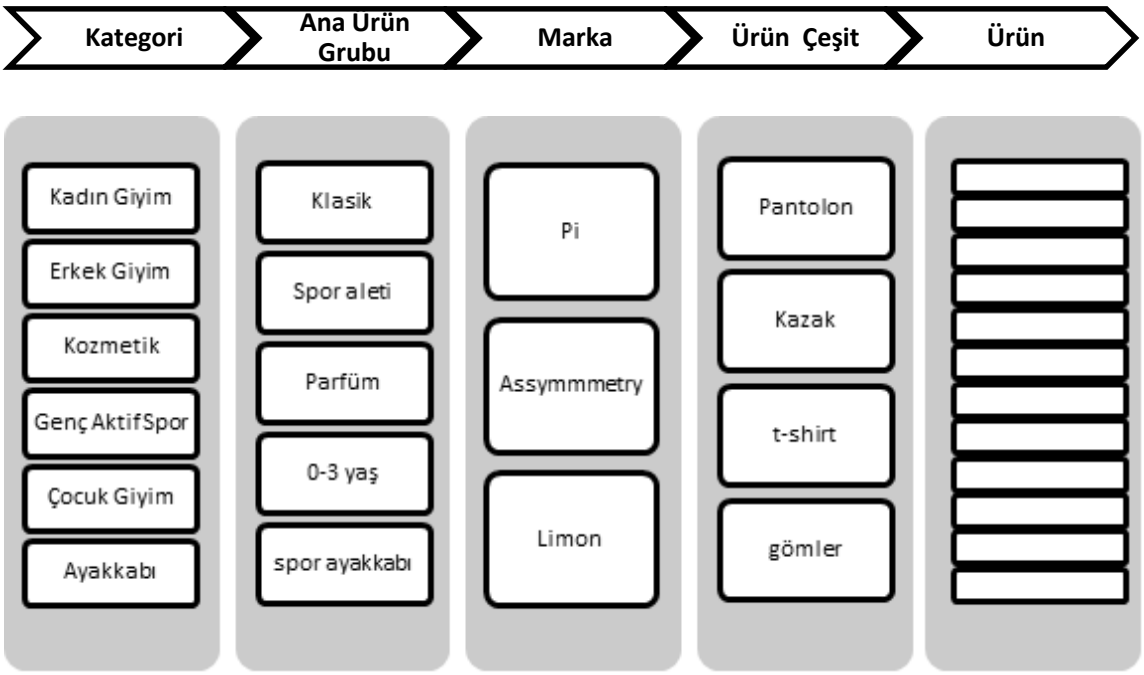
ilk iki veri minimum seviyelerde veri madenciliği süreçleri için olması gereken veri tipleridir.

7.1 Ürün Verisi

Ürün verileri veri madenciliği için en temel verilerdir. Bu veriler standart bir kod (UPC-universal product code) veya perakendeciler tarafından kullanılan SKU (stock-keeping

units) formundadır. Burada ürün adı, marka ve fiyat bilgileri gibi kırılımlar oluşmaktadır. Birçok perakendeci ürün bilgilerini tanımlayıcı bir hiyerarşi yapısında yönetmektedir.

Bu hiyerarşik yapıda en üst seviye ürün grupları (yiyecek, alkol, ev ürünleri, kozmetik)ürün alt gruplarına (alkol-bira, şarap) bölünmektedir. Her bir alt grup daha sonra ürün kalemlerine (A-birası-2' li paket)ayrılmaktadır. İşletmeye göre kırılım sayıları değişim gösterebilmektedir. Bizim örneğimizde ürün veri yapısı:



Şekil 7.1 Ürün verisi kırılımları

7.2 İşlemsel Verileri

Birçok perakende firması elektronik olarak satış noktalarından toplanan mağaza bilgilerini toplamaktadır. İşlemsel veriler satış anında oluşan işlem kayıtlarını:

- İşlem tarih ve zamanı
- Lokasyon bilgisi
- UPC veya diğer ürün kodları
- Satılan ürün adedi
- Ödenen miktar
- Ödeme tipi (nakit, kredi kartı, debit kartlar)

İşlemsel veriler ile veri madenciliği süreçleri uygulanabilmesine karşın, işlemlerin tekil tanımlı müşterilere bağlanabilmesi, işletmeye büyük avantajlar sağlayabilmektedir. Bu durumlarda tekil tanımlı bir numara ile aynı müşteriyi çoklu işlemlerde takip edebilme ve zaman içinde müşteri profilleri oluşturabilme avantajı elde edilmektedir. Bu nedenle müşteriyi tanımlayan bir tekil numara listeye eklenmelidir.

Çoğunlukla bu tekil tanımlayıcı sadakat kart uygulamaları ile elde edilmektedir. Bu tekil tanımlayıcı ile çoklu işlem verilerinin yanı sıra, diğer veri kaynakları ile de ilişkilendirilebilir. Sadakat kart programlarının temel amacı, müşterilere bir takım ürünler ve farklı ödüllerle değişebilir bonus puanlar kazandırarak, onların sadık kalmalarını sağlamaktır.

7.3 Müşteri Demografik Verileri

Demografik veriler bireysel müşterilerle ilgili tanımlayıcı verilerdir.

- Yaş
- Cinsiyet
- Eğitim durumu
- Ev adresi
- Medeni hal

Eğer müşteri sadakat kart programının üyesiyse bu bilgilerin birçoğu bulunmaktadır ve bulunmalıdır. Birçok müşteri bu bilgileri paylaşmakta isteksizde olsa, bu bilgilerle dokunulur bir faydanın sunulması müşterileri ön yargılarını kaldırabilmektedir.

Demografik bilgilerin kullanımı sırasında önemli bir nokta bu bilgilerin doğru olup olmadığının kontrol edilmesidir.

7.4 Müşteri İlişkileri Verileri

Müşteri ilişkileri verileri, müşteri ilişkileri yönetiminin direk bir sonucu olarak ortaya çıkan kurum içinde türetilen verilerdir.

- Müşteri hedefli promosyonlar
- Önceden tanımlanmış müşteri segmentleri (veri madenciliği yada bir takım işletme kuralları ile tanımlanmış olabilir)
- Ürün tercihleri

- Kanal kullanımı
- Şikayetler

7.5 Ek Veriler

Ek veriler müşteri davranışlarını anlamaya yönelik destek olabilmektedir. Bu değişkenler kuruma göre değişebilir, ancak şöyle sıralanabilir:

- Hava şartları
- Rakip firmaların konumu ve mevcut promosyonları
- Genel ekonomik veriler, gayri safi milli hasıla
- Ekonomik krizler
- Kalitatif araştırma sonuçları

Bu listenin sonu olmamakla beraber, bu tür veriler veri madenciliğinde belirli süreçlerde kullanılması önemli olabilmektedir.

Perakende firmalarında başarılı olan ve geliştirilen iki veri modeli tanımlanmaktadır. İlk model, doğrudan satış noktalarından elde edilen işlem verilerinden çekilen bilgileri ve bu bilgileri hiyerarşik ürün yapısını kullanarak ürün datası ile birleştirilmesi ile elde edilmektedir. İkinci model ise birinci modeli tekil müşteri tanımlayıcısı kullanarak, çoklu işlemlerin birleştirilerek daha gelişmiş bir model oluşturulmasıdır.

7.6 İşlem Bazında Veri Modeli

Kuşkusuz tekil bir müşteri numarası ile işlem verileri veri madenciliğini oldukça değerli kılmakta ve bu bilgi veri madenciliğinin değerini anlamlı bir şekilde arttırmaktadır. İşlem bazlı veriler veri madenciliğinde, ürün ilişkilerini inceleyen Sepet analizinde kullanılabilmektedir.

Müşteri davranışlarını tanımlayabilmek adına gereken minimum bilgiler:

- İşlem numarası
- İşlem tarihi ve zamanı
- Satın alınan ürün (UPC)
- Ürün Fiyatı
- Satın alınan ürün adedi

- Ürün kodunu ürün ismiyle, alt grup kodlarını alt grup isimleriyle, grup kodlarını da grup ismiyle eşleştiren tablolar
- Ürün sınıflandırması ürün kodlarının alt gruplara ve gruplara bağlanması
- Her bir işlem kaydı aşağıdaki alanları içermektedir:
- İşlem numarası
- İşlem tarihi ve zamanı
- İşlemin toplam cirosu
- İşlemdeki toplam ürün adedi
- Her bir üründe nispi harcama
- Normalize edilmiş nispi harcama (NNH-NRS)
- Alt grup bazında nispi harcama
- Her bir alt grupta NNH
- Her bir ürün grubunda nispi harcama
- Her bir grupta NNH

Nispi harcama, her bir grupta harcama oranları ile tanımlanmaktadır. Örneğin bir işlemde, bir müşteri 4 ürün için 5 TL, 20 TL, 25 TL ve 50 TL harcamış olsa, her bir 4 ürünün nispi harcamaları 0.05, 0.2, 0.25 ve 0.5 olacaktır. İlk iki ürün aynı alt grupta, diğer iki ürün ise başka bir alt grupta yer alıyorsa, nispi harcamalar 0.25 ve 0.75 olacaktır.

Normalize nispi değerler, ürünün tüm satışlara kıyaslamalı anlamlı harcamaları hesaplamaya çalışır. Örneğin A ürününün nispi harcama oranı %50, fakat A ürününün tüm satışlar içerisindeki oranı da %50 ise müşteri normal satın alım davranışı gösteriyor demektir. A ürünü için normalize nispi harcaması 1 olacaktır. Eğer A ürünü tüm satışların %10 kadar olsaydı, müşteri A ürününü genele göre beş kat daha fazla satın almıştır. Bu nedenle NNH 5 değerini alacaktır.

7.7 Uygulama Verisi

Uygulamanın gerçekleşeceği çalışmada kullanılacak olan veriler bir perakende firmasına ait olup, satış bilgileri (fatura) ile sadakat kart bilgilerinin birleştirilmesi ile elde edilen bir veri yapısına sahiptir.

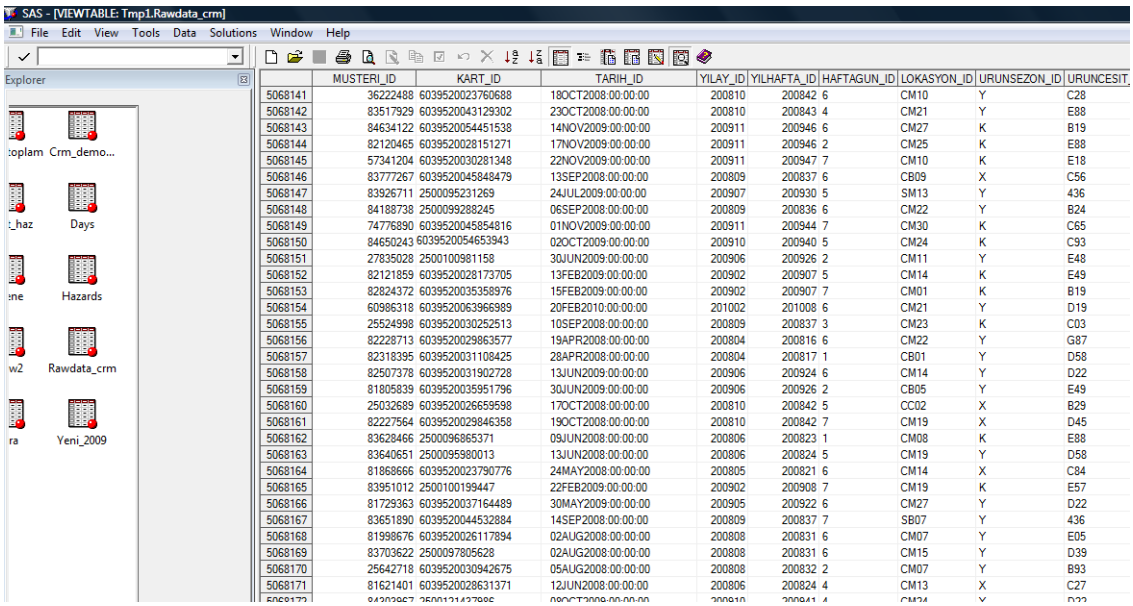
Firma müşterileri alışverişlerini gerçekleştirirken sahip oldukları sadakat kartının faydası gereği, alışverişleri esnasında kullanmaları ile fatura bilgileri ile müşteri bilgileri eşleşebilmektedir.

Sadakat kart kullanım oranının %50 üzerinde olması, yapılacak çalışmalar için değerli ve anlamlı bilgiler sunabilmektedir. Müşterilerin izinli pazarlama kapsamında iletmiş oldukları bilgiler sonucunda, müşteri sadakat sistemine dahil olmaktadır. Sadakat kartı, müşterinin alışveriş alışkanlıkları belirlenebilmekle beraber, sadakat sistemi üzerinden aksiyon alınmasını ve özelleştirilmiş faydaların sunulmasına imkan tanımaktadır.

Önem verilmesi gereken noktalar;

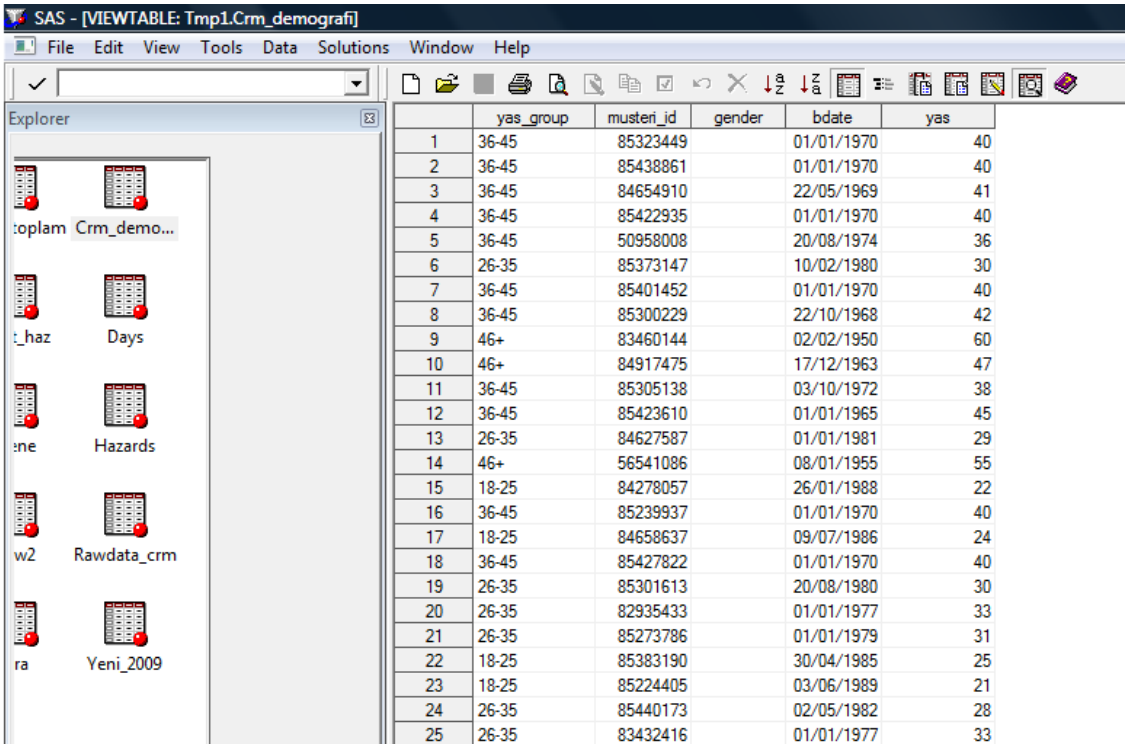
- Doğru analiz ve sonuçlar için, her bir müşterinin tek bir kartı olmalı
- Müşterinin tanımlı tek kartı olmasa dahi, kartların tek bir tanımlamada toplanması gerekmektedir
- Kartın kullanımı, sadece ve sadece kartın sahibine ait olmalıdır
- Her bir müşterinin kart kullanımının yüksek oranlarda olması gerekmektedir

Çizelge 7.1 Örnek Müşteri Transaction Tablosu (SAS library)



	MUSTERI_ID	KART_ID	TARIH_ID	YILAY_ID	YILHAFTA_ID	HAFTAGUN_ID	LOKASYON_ID	URUNSEZON_ID	URUNCESIT
5068141	36222488	6039520023760688	18OCT2008:00:00:00	200810	200842	6	CM10	Y	C28
5068142	83517929	6039520043129302	23OCT2008:00:00:00	200810	200843	4	CM21	Y	E88
5068143	84634122	6039520054451538	14NOV2009:00:00:00	200911	200946	6	CM27	K	B19
5068144	82120465	6039520028151271	17NOV2009:00:00:00	200911	200946	2	CM25	K	E88
5068145	57341204	6039520030281348	22NOV2009:00:00:00	200911	200947	7	CM10	K	E18
5068146	83777267	6039520045848479	13SEP2008:00:00:00	200809	200837	6	CB09	X	C56
5068147	83926711	2500095231269	24JUL2009:00:00:00	200907	200930	5	SM13	Y	436
5068148	84188738	2500099288245	06SEP2008:00:00:00	200809	200836	6	CM22	Y	B24
5068149	74776890	6039520045854816	01NOV2009:00:00:00	200911	200944	7	CM30	K	C65
5068150	84650243	6039520054653943	02OCT2009:00:00:00	200910	200940	5	CM24	K	C93
5068151	27835028	2500100981158	30JUN2009:00:00:00	200906	200926	2	CM11	Y	E48
5068152	82121859	6039520028173705	13FEB2009:00:00:00	200902	200907	5	CM14	K	E49
5068153	82824372	6039520035358976	15FEB2009:00:00:00	200902	200907	7	CM01	K	B19
5068154	60986318	6039520063966989	20FEB2010:00:00:00	201002	201008	6	CM21	Y	D19
5068155	25524998	6039520030252513	10SEP2008:00:00:00	200809	200837	3	CM23	K	C03
5068156	82228713	6039520029863577	19APR2008:00:00:00	200804	200816	6	CM22	Y	G87
5068157	82318395	6039520031108425	28APR2008:00:00:00	200804	200817	1	CB01	Y	D58
5068158	82507378	6039520031902728	13JUN2009:00:00:00	200906	200924	6	CM14	Y	D22
5068159	81805839	6039520035951796	30JUN2009:00:00:00	200906	200926	2	CB05	Y	E49
5068160	25032689	6039520026659598	17OCT2008:00:00:00	200810	200842	5	CC02	X	B29
5068161	82227564	6039520029846358	19OCT2008:00:00:00	200810	200842	7	CM19	X	D45
5068162	83628466	2500096865371	09JUN2008:00:00:00	200806	200823	1	CM08	K	E88
5068163	83640651	2500095980013	13JUN2008:00:00:00	200806	200824	5	CM19	Y	D58
5068164	81868666	6039520023790776	24MAY2008:00:00:00	200805	200821	6	CM14	X	C84
5068165	83951012	2500100199447	22FEB2009:00:00:00	200902	200908	7	CM19	K	E57
5068166	81729363	6039520037164489	30MAY2009:00:00:00	200905	200922	6	CM27	Y	D22
5068167	83651890	6039520044532884	14SEP2008:00:00:00	200809	200837	7	SB07	Y	436
5068168	81998676	6039520026117894	02AUG2008:00:00:00	200808	200831	6	CM07	Y	E05
5068169	83703622	2500097805628	02AUG2008:00:00:00	200808	200831	6	CM15	Y	D39
5068170	25642718	6039520030942675	05AUG2008:00:00:00	200808	200832	2	CM07	Y	B93
5068171	81621401	6039520028631371	12JUN2008:00:00:00	200806	200824	4	CM13	X	C27

Çizelge 7.2 Müşteri Demografik Verileri



	yas_group	musteri_id	gender	bdate	yas
1	36-45	85323449		01/01/1970	40
2	36-45	85438861		01/01/1970	40
3	36-45	84654910		22/05/1969	41
4	36-45	85422935		01/01/1970	40
5	36-45	50958008		20/08/1974	36
6	26-35	85373147		10/02/1980	30
7	36-45	85401452		01/01/1970	40
8	36-45	85300229		22/10/1968	42
9	46+	83460144		02/02/1950	60
10	46+	84917475		17/12/1963	47
11	36-45	85305138		03/10/1972	38
12	36-45	85423610		01/01/1965	45
13	26-35	84627587		01/01/1981	29
14	46+	56541086		08/01/1955	55
15	18-25	84278057		26/01/1988	22
16	36-45	85239937		01/01/1970	40
17	18-25	84658637		09/07/1986	24
18	36-45	85427822		01/01/1970	40
19	26-35	85301613		20/08/1980	30
20	26-35	82935433		01/01/1977	33
21	26-35	85273786		01/01/1979	31
22	18-25	85383190		30/04/1985	25
23	18-25	85224405		03/06/1989	21
24	26-35	85440173		02/05/1982	28
25	26-35	83432416		01/01/1977	33

7.8 Müşteri Bazında Veri Modeli

Eğer müşteri segmentasyonu gerçekleştiriliyorsa, tüm veriler müşteri seviyesinde toplanmalıdır.

Müşteri modeli oluşturabilmek için bahsedilen işlemsel verilerin belirli bir zaman aralığında her bir tekil müşteri bilgileri ile üretilen veri şu değişkenleri içermektedir:

- Tekil müşteri tanımlayan bir numara
- Müşterinin toplam harcaması
- Her müşteri için işlem sayısı (sıklık)
- İşlemler arasındaki ortalama süre (işlem aralığı) (1)
- İşlem aralığı değişimi (2)
- Müşteri istikrar indeksi (2)/(1)
- Son ziyaret sonrası geçen süre (sonzamandardalık)
- Her işlemde ortalama farklı ürün sayısı
- Her ürün için nispi harcama
- Her bir ürün için NNH (NRS-Normalized Ratio Spent)
- Her ürün alt grup için nispi harcama oranları
- Her ürün alt grup için NNH

- Her ürün grubu için nispi harcama
- Her ürün grubu için NNH

Gerçek bir perakende firmasına ait değişken yapısı (Çizelge 7.1)' deki tabloda görülebilmektedir. Değişkenler çok katlı mağazacılık alanında işlev gören ticari bir şirkete aittir.

Çizelge 7.3 Gerçek bir perakende firmasına ait verilerden oluşturulan datamart yapısı

Kod	**Değişken**
DHT	Analiz dönemi boyunca toplam harcama miktarı
DINT	Analiz dönemi boyunca toplam işlem miktarı
DFNT	Analiz dönemi boyunca toplam fatura adedi
DMHO	Analiz dönemi boyunca aylık ortalama harcama
DYHO	Analiz dönemi boyunca yıllık ortalama harcama
DI_H	Analiz dönemi boyunca işlem başına harcama miktarı
DF_H	Analiz dönemi boyunca fatura başına harcama miktarı
DMFNO	Analiz dönemi boyunca aylık ortalama fatura sayısı
DYFNO	Analiz dönemi boyunca yıllık ortalama fatura sayısı
DFF1	Tüm analiz dönemi alışveriş sıklığı
DFP1	Tüm analiz dönemi alışveriş periyodu
SFIF	İlk işlem tarihinden itibaren geçen süre (GÜN)
SFIL	Son işlem tarihinden itibaren geçen süre (GÜN)
SA	Aktivite süresi (GÜN)
ASFIF	İlk işlem tarihinden itibaren geçen süre (AY)
SE	Etkin Müşteri Olma Süresi - Ay bazlı
SEG	Etkin Müşteri Olma Süresi - Gün bazlı
SE2	Etkin Müşteri Olma Süresi - Ay bazlı
SEG2	Etkin Müşteri Olma Süresi - Gün bazlı
MN	Alışveriş Yapılan Farklı Ay Sayısı
LXHT	X Mağazası Toplam Harcama
LXFT	X Mağazası Toplam Fatura
LXIT	X Mağazası Toplam İşlem
LXHR	X Mağazası Harcama oranı
LXIR	X Mağazası Alışveriş İşlem oranı
LXFR	X Mağazası Fatura oranı
vl	Tercih edilen farklı lokasyon sayısı
BN	Tercih Edilen Farklı Marka Sayısı
CN	Tercih Edilen Farklı Kategori Sayısı
UGN	Tercih Edilen Farklı Ürün Grubu Sayısı
CXHT	X Kategorisi toplam harcama miktarı
CXIT	X Kategorisi toplam işlem miktarı
CXFT	X Kategorisi toplam fatura sayısı

Çizelge 7.3 Gerçek bir perakende firmasına ait verilerden oluşturulan datamart yapısı…(devamı)

CXF_H	X Kategorisi fatura başına harcama
CXI_H	X Kategorisi işlem başına harcama
CXMHO	X Kategorisi aylık ortalama harcama
CXMIO	X Kategorisi aylık ortalama işlem
CXMFO	X Kategorisi aylık ortalama fatura
CXBN	X Kategorisi tercih edilen farklı marka sayısı
CXUGN	X Kategorisi tercih edilen farklı ürün grubu sayısı
CXUN	X Kategorisi tercih edilen farklı ürün çeşidi sayısı
DCXHR	Analiz dönemi boyunca X kategorisi harcama oranı
DCXIR	Analiz dönemi boyunca X kategorisi alışveriş oranı

BÖLÜM 8

MÜŞTERİ SEGMENTASYONU

İdeal dünyada kurumlar her bir müşterisine bireysel hizmet ederek müşteri memnuniyeti ve karlılık sağlamayı istemektedir. Günümüz CRM programlarının da amacı budur. Gerçek dünyada ise bire bir etkileşim ekonomik olarak mümkün olamamaktadır. Bu nedenle işletmeler müşterilerini onlara bire bir pazarlama anlayışını işletmenin zaman ve maliyeti kapsamında belirli gruplara bölmektedirler.

Bazı tanımlı işletme tanımlarının aksine, veri madenciliğini kullanarak müşteri segmentleri veriden elde edilmektedir. Belirli müşteri davranışlarının (değişkenlerinin) veritabanından direk alınması veya başka bilgi kaynaklarından alınan bilgilerle birleştirilmesi ile elde edilen veri içerisinde müşterilerin doğal gruplarına ayrılma sürecidir.

Segmentasyon farklılaştırılmış pazarlama stratejileri geliştirme amaçlı müşterileri karakteristik yapıları veya davranışları açısından benzerlik veya farklılıklarına göre gruplama yöntemidir. İşletme amaçlarının tanımıyla başlar, segmentler için farklılaşmış pazarlama stratejilerinin sunulması ile biter. Farklı kriter ve davranışlara göre bir çok segmentasyon yöntemi bulunmaktadır. Müşteriler işletmeye olan değerlerine göre, sosyo-demografik ve hayat evreleri bilgilerine göre, davranışlarına göre, istek ve ihtiyaçlarına göre, sadakat ölçülerine göre gruplandırılabilirler. Kullanılacak segmentasyon yöntemi işletmenin amaçlarına göre belirlenmelidir.

8.1 Segmentasyonun Sonuçları

- Yeni pazar fırsatlarını belirlemeye yönelik **müşterileri, tüketicileri anlayabilmek**

- **Yeni ürün veya hizmet geliştirebilmek** ve kitlesel pazar yerine **segment karakteristikleri ile örtüşen ürün geliştirmek**
- **İstek ve ihtiyaçları tanımlanan her bir segmente göre**, mevcut müşterilere **kişiselleştirilmiş ürün tasarlamak**
- **Ödül ve teşvik edici fırsatlar sunmak**
- **Uygun reklam, iletişim dili ve kanal seçebilmek**
- **Uygun satış ve hizmet kanalı seçmek**
- **Ürün faydasını ve marka imajını belirleyerek,** her bir segment karakteristiğine göre **iletişim sağlamak**
- Her segmentin önemine göre **müşteri servisinde farklılaşmak**
- Her bir segmentin potansiyel geri dönüşlerine göre **kaynak kullanımı** ve **kaynak tahsisi**
- **Önceliklendirme**, her bir segmentin önemi ve değerine göre elde tutma ve geliştirme programları

8.2 Başarılı Bir Segmentasyonun Özellikleri

- Kurum tarafından belirlenmiş işletme amaçlarına hitap etmeli.
- Tekil karakteristik ve ihtiyaçlara sahip farklılaşmış segment tanımlarının belirlenmesi gerekmektedir.
- Her bir bireysel bir müşteri bir segmente doğru bir şekilde atanmış olmalıdır.
- **Belirlenmiş segment yapıları kar/fayda için fırsatlar sağlamalıdır.**
- **Segmentler anlaşılır profil ve karakteristik yapılar sergilemelidir.**
- Kurum ortaya çıkan segmentler sayesinde etkili pazarlama aksiyonları tasarlayabilmelidir. İşletme faaliyetlerine uygulanabilir ve aksiyon alınabilir olmalıdır.

BÖLÜM 9

SEGMENTASYON MODELLERİ

Birçok kritere göre müşteri segmentasyon modelleri uygulanmaktadır. Her bir işletmenin gereksinimleri ve amaçlarına göre segmentasyon modelleri tercih edilmektedir. Segmentasyon öncesinde işletme amaçları ve başarı kriteri mutlaka tanımlanmalıdır.

- Değer bazlı segmentasyon
- Davranış bazlı segmentasyon
- Eğilim bazlı segmentasyon
- Sadakat bazlı segmentasyon
- Sosyo-demografik ve Hayat-evreleri bazlı segmentasyon
- İstek ve ihtiyaç bazlı segmentasyon

9.1 Değer Temelli Segmentasyon

Değer temelli segmentasyon müşterileri işletmeye olan değerleri açısından bölümlendirme yöntemidir. Bu işletmeye zamana karşı değer değişimlerinin takibini sağlamaktadır. Bu şekilde işletmeler aşağı göçleri önlemiş, yukarı göçleri de destekleyebilecek bilgiye sahip olabilmektedirler.

Değer bazlı segmentasyon modeli karmaşık veri madenciliği modellerine ihtiyaç duyulmadan da kolay hesaplamalar ile elde edilebilmektedir. Müşterilerin değerlerine göre sıralanmaları ve sıklık derecelerine göre değer grupları belirlenebilir:

- **Gold**: En üst %20' lik dilimde yeralan yüksek değer grubu
- **Silver**: Sonraki %30' luk dilimde yeralan ikinci yüksek değer grubu

- **Bronz**: Alt %50' lik dilim düşük değer grubuna sahip müşteriler

9.2 Davranışsal Segmentasyon

Oldukça yararlı ve etkili bir segmentasyon yöntemidir. Veri ambarı ve datamart yapılarından elde edilebilecek müşteri ürün sahipliği ve kullanım verilerine (işlem hacmi/sıklığı/tipi; ödeme tarihçesi) ihtiyaç duyulmaktadır. Müşteriler belirli kullanım alışkanlıklarına ve davranışlarına göre gruplandırılmaktadır. Yeni ve kişiselleştirilmiş ürün tekliflerinin geliştirilmesi ve sadakat şemalarının tasarlanmasında kullanılmaktadır.

Örneğin bir banka için genel davranış segmentleri:

- **Gelecek yatırımcıları:** Sigorta veya yatırım ürünleri gibi kalemlerde işlem yapan müşteriler
- **Sık gezen:** Bir çok farklı kanal ve noktadan sık ve çeşit ürün harcaması yapan müşteri grubu
- **Kredici:** Ağırlıklı olarak ev ve ihtiyaç kredisi işlemi yapan müşteriler

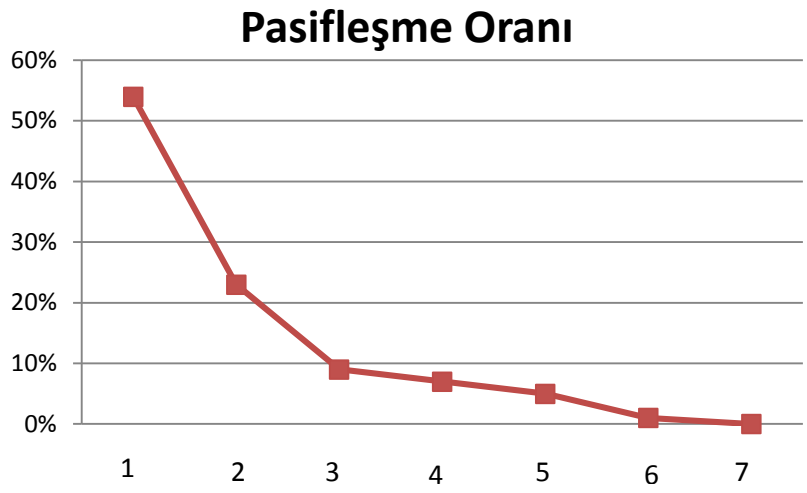
Bir telekomünikasyon şirketinde ise:

- **Roaming' ciler:** Mobil telefonunu yurt dışı aramalarında kullanan, yurt dışından arama yapan müşteriler
- **Süperstarlar:** Neredeyse mevcut tüm servisleri kullanan (internet, MMS, 3G), yeni çıkan servislere hemen adapte olan müşteriler
- **Klasik kullanıcılar:** ortalama konuşma yapan ve sms ağırlıklı iletişim yapan müşteriler gibi davranış bazlı segmentasyon olarak nitelendirilebilecek farklı sektör uygulamalarıdır.

9.3 Eğilim Bazlı Segmentasyon

Eğilim bazlı segmentasyon modelleri çapraz satış ve müşteri kaybı analizlerinin sonuçlarından elde edilen segmentasyon tipidir. Müşteri kaybı (churn) modelinde her müşterinin müşteri eğilim skorları hesaplanmaktadır. Skorlar ayrılma olasılıklarını belirleyerek, pazarlamacılara müşterilerini kaybetme risklerine göre sıralamalarında yardımcı olmaktadır. Atanan bu skorlar sayesinde belirli eşik değerlerde, müşteriler düşük, orta ve yüksek risk gruplarına ayrılabilmektedir. Risk grupları, diğer segmentasyon yöntemleri ile birlikte müşteriyi anlamada daha zengin bilgi

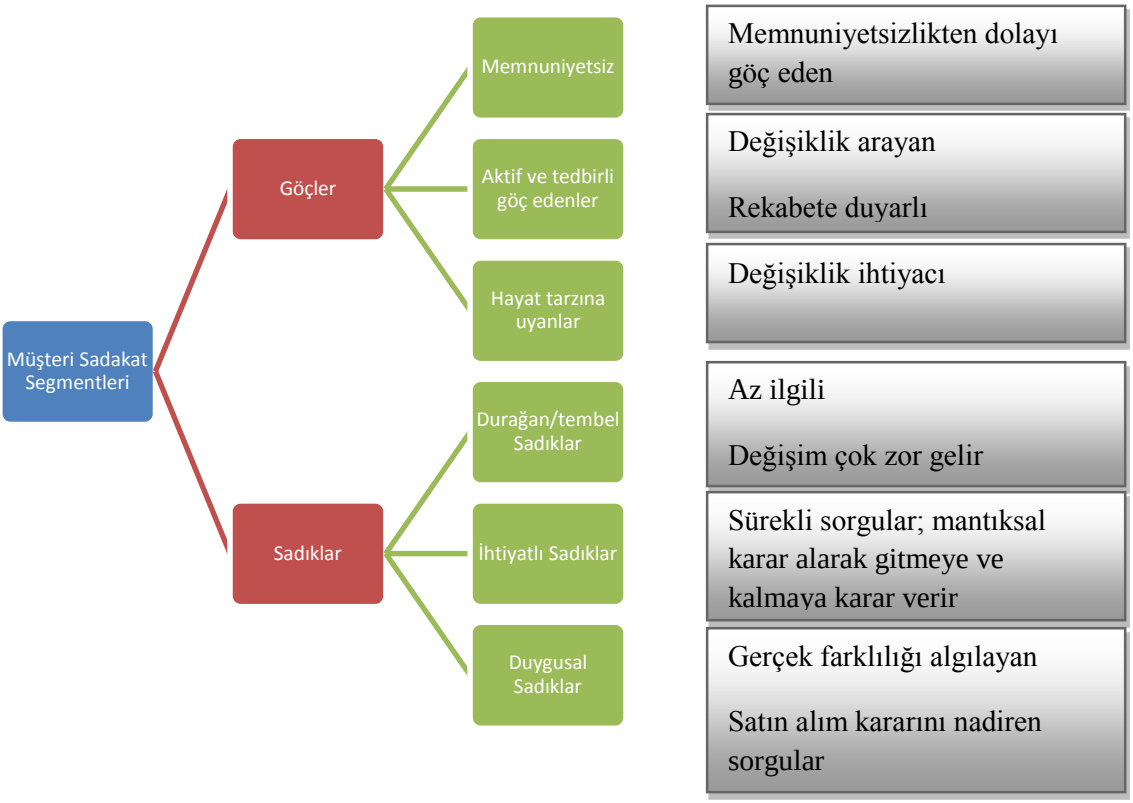
sunabilmektedir. (Şekil 9.1)' de churn skoru düşük olan müşterilerin, 7 değer grubundaki oranları gösterilmektedir. Değer grubunda 7 değeri en düşük, 1 en yüksek değer grubu olmak üzere churn skoru düşük olan müşterilerin oranları görülmektedir. 7. Değer grubunda yer alan müşterilerin %60' ının churn skorları düşük olan grupta yer almaktadır. 1. değer grubunda ise neredeyse hiç churn skoru düşük müşteri bulunmamaktadır.



Şekil 9.1 Değer bazlı segmentasyon ile churn skoru düşük müşterilerin dağılımı, riskde değer segmentasyonu

9.4 Sadakat Segmentasyonu

Bu segmentasyon modelinde sadakat durumlarına göre müşterileri belirli gruplar belirlenir ve bu müşterileri göç eden müşterilerden ayrılır. Bu segmentler bazı işletme kurallarına göre ayrılabileceği gibi veritabanı bilgisiyle kümeleme algoritmaları kullanılabilmektedir.



Şekil 9.2 Sadakat Segmentleri

Bir GSM operatörü için örnek sadakat temelli segmentasyon:

Memnuniyetsizler: Arama merkezine yapılan şikayetler

Aktif göçenler: Her zamanen uygun ürün veya programla beraber, sürekli ürün veya program değişikliği. Genellikle yeni müşteri

Yaşam tarzına uyanlar: Yaşam evrelerindeki değişiklikleri gösteren kullanım alışkanlıkları değişimleri

Durağan sadıklar: Kullanım profillerine uygun olmamasına karşın bile sürekli aynı program ve ürünle

İhtiyatlı sadıklar: Her zaman en uygun ürün veya programla, normal seviyelerde program veya ürün değişikliği, genellikle yaşlı müşteriler

Duygusal sadıklar: Sadıklar ama hem durağan, hem de ihtiyatlı sadıklardan farklılar.

9.5 Sosyo-Demografik ve Yaşam-Evresi Bazlı Segmentasyon Modeli

Demografik segmentasyon müşterilerin demografik yapılarına göre gruplanmaları sonucu elde edilmektedir. müşteri demografik bilgileri onların ihtiyaçlarına, önceliklerine ve tüketme/kullanma alışkanlıklarına kuvvetli bir şekilde etki etmektedir. Buna karşın aynı demografik yapıda olan insanların farklı ihtiyaç ve istekleri olmaktadır. genellikle bu segmentler şu demografik özellikler ile gruplandırılır:

- Cinsiyet
- Sosyal konum
- Eğitim
- Yaş
- Medeni hal
- Ev sahipliği
- Aile geliri
- Aile bireyi sayısı
- Meslek
- Çocukların yaşları

Aile hayat döngüsü segmentasyonu müşteri yaşı, medeni hali, çocuk sayısı ve yaşları ile tanımlanmaktadır. Genel bir aile hayat döngüsü segmentleri:

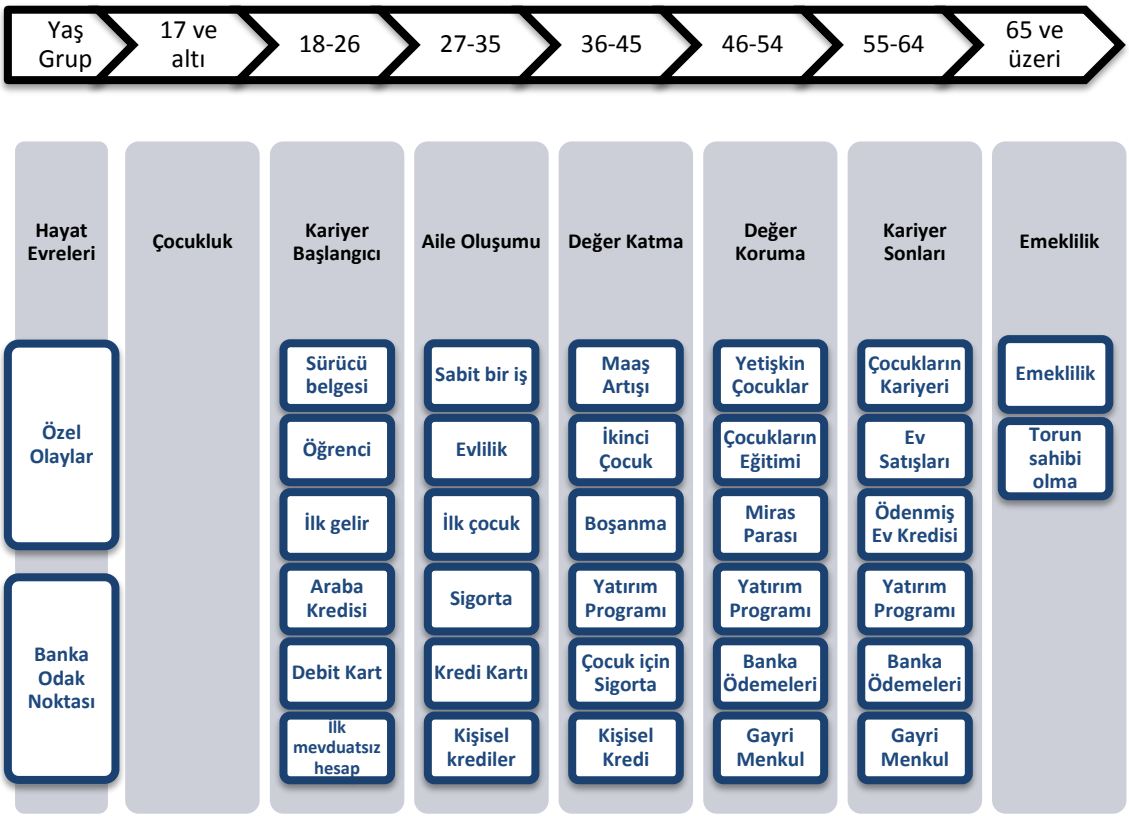
- Genç, bekarlar
- Genç evliler, henüz çocuk yok
- Genç aileler, 5 yaşından ufak çocuk sahibi
- Genişleyen aileler, 5 yaşından büyük çocuk sahibi
- Çocuklu aileler
- Emekli daha yaşlı kişiler

Demografik segmentasyon (yada bilgiler), davranış, değer ve eğilim bazlı segmentasyon modellerini geliştiren bilgiler sunmaktadır.

Hayat evreleri kişinin önceliklerini ve hayattaki asıl önemli olayları tanımlamaktadır. Bu ilk çocuğun doğumu, anlamlı bir gelir artışı gibi nedenler olabilmektedir. Hayat evreleri özellikle ihtiyaçlarına hitap edebilecek ürün ve hizmet kampanyalarının oluşmasını sağlayabilmektedir. Elde etmesi kolay olmayan hayat-evresine ilişkin veriler müşteri

tüketim davranışları ile birleştirildiğinde çok değerli fırsatlar sunabilmektedir. Örneğin bir banka için çocuk doğumu ev kredi ihtiyacı ile yüksek derecede ilişkilidir. Diğer taraftan yıllık gelirdeki anlamlı artış ise yatırım ve sigorta ihtiyaçları getirebilir. Çoğunlukla kurumun veritabanlarında mevcut olmayan hayat-evreleri olayları müşteri yaş grubu kırılımında tahmin edilmektedir.

Banka ve finans alanından bir örnek (Şekil 9.3)' de gösterilmektedir.



Şekil 9.3 Müşteri yaşam evreleri banka örneği

9.6 İhtiyaç Bazlı Segmentasyon

İhtiyaç bazlı segmentasyon müşteri ihtiyaçlarını, isteklerini, önceliklerini ve algılarını incelemede kullanılmaktadır. İlgili bilgi Pazar araştırmalarından ve anket sonuçlarından toplanan kümeleme uygulamaları ile elde edilebilmektedir.

Veri madenciliği çerçevesinde, ihtiyaç bazlı segmentasyon davranış ve değer bazlı segment profillerini zenginleştirip, kalitatif karakteristiklerin ortaya çıkmasını sağlar ve:

- Yeni ürün tasarım ve geliştirmesini
- Her bir segment için marka imajı ve ürün özelliklerinin iletişiminin gerçekleştirilmesini

- Segmentlere bire bir uyan iletişim mesajları iletebilmeyi
- Yeni promosyonları desteklemektedir.

BÖLÜM 10

SGP (Sonzamandardalık, Geliş Sıklığı, Parasal) – RFM ANALİZİ

SGP oldukça yaygın kullanıma sahip müşteri harcama alışkanlıklarının anlaşılmasına yönelik gerçekleşen bir yaklaşımdır. Özellikle perakende alanında oldukça popüler bir yöntemdir. **Sonzamandardalık, Geliş Sıklığı, Parasal** olmak üzere üç anahtar performans göstergesi kurum ile müşteri arasındaki ilişkinin boyutuna ilişkin özet bilgi sunabilmektedir. **Sonzamandardalık** müşterinin son işlem tarihinden güncel zamana geçen süreyi, **geliş sıklığı** müşterinin satın alma işlem sayısını, **parasal** ise harcamaların değerini göstermektedir. Bu göstergeler tekil müşteri tanımlayıcısı seviyesindeki mevcut işlemsel veri modelinden hesaplanabilmektedir.

Bu model müşteri değer skorlanmasında da kullanılmaktadır. Dolayısı ile bu yöntemin müşteriyi tanımlamada yardımcı bir rolü bulunmaktadır.

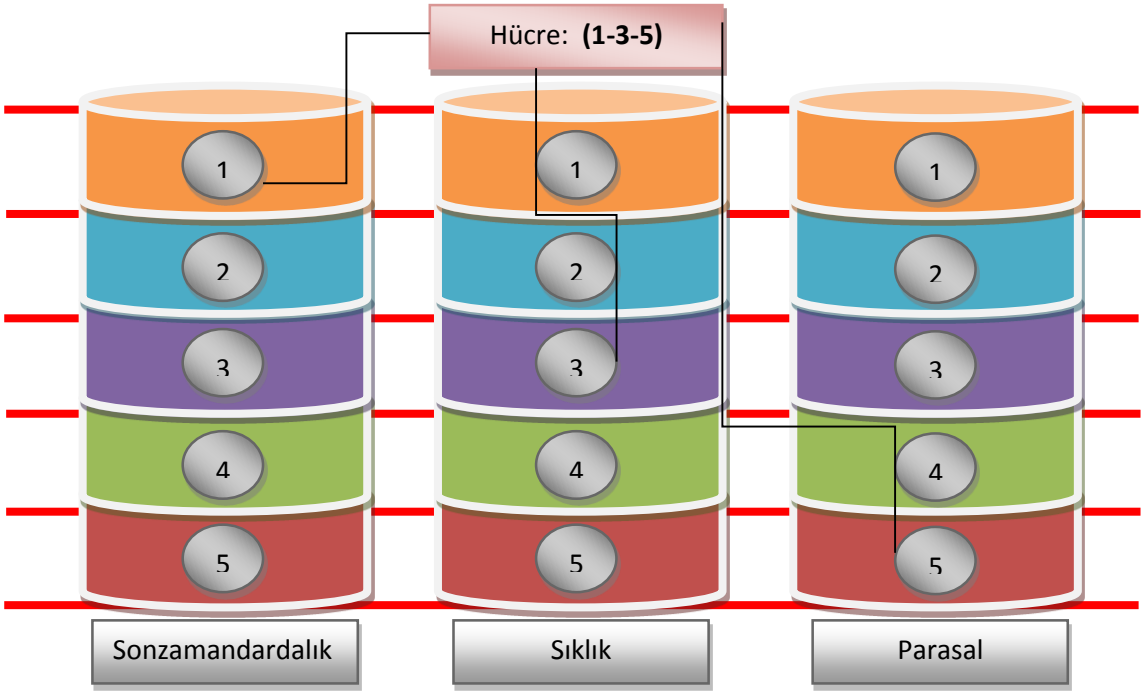
Sonzamandardalık: Son işlem veya son alışveriş ziyaretinden itibaren geçen süre (gün/ay/yıl)

Sıklık: Belirli bir dönemdeki toplam alışveriş işlem sayısı veya alışveriş ziyaret sayısı. Alternatif olarak birim zamandaki işlem sayısı, örneğin aylık ortalama işlem sayısı.

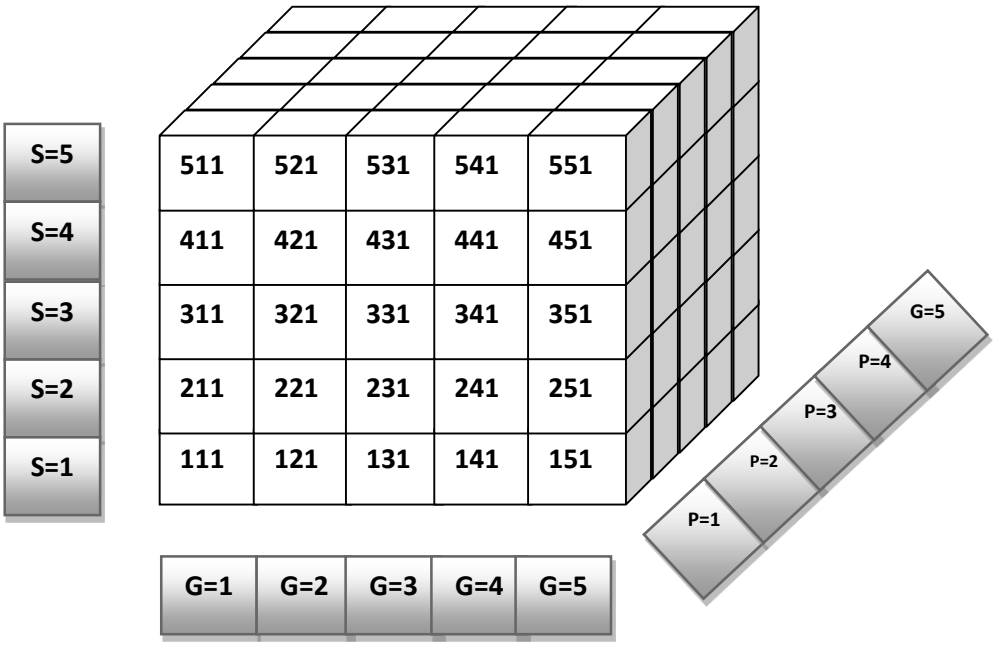
Parasal: İncelenen periyottaki toplam alışveriş değeri yada belirli birim zamandaki ortalama değer.

SGP göstergelerinin oluşumu basit veri yönetim işlemleri gerektirmektedir. Her bir performans göstergesinin anlamlı bir skorda birleştirilmesi gerekmektedir. Bu performans indeksleri aynı zamanda Davranışsal Segmentasyon sürecinde de girdi olarak kullanılmaktadır. Genellikle SGP müşterileri eşit kitlelerde veya istatistiki sıklık

derecelerine (kantil) göre gruplandırılmaktadırlar (Değer bazlı segmentasyon gibi). Bu gruplama bağımsız olarak üç performans göstergelerine uygulanmaktadır. Müşterilerin ilgili ölçülerde sıralanması ve eşit boyutlardaki sınıflarca gruplanması işlemidir. Bu gruplama yöntemlerinde farklı yaklaşımlarda kullanılmaktadır. Sonuç olarak SGP değerleri, sıralı ölçümlü skorlara dönüştürülmektedir. Kantil gruplamasında SGP ölçüleri 1 den 5' e sıra skorlarına dönüştürülmektedir. 1. grup en düşük dilimde yeralan %20' lik kısmı, 5. grup ise en yüksek %20' lik dilimi ifade etmektedir.



Şekil 10.1 hücre atamaları



Şekil 10.2 Kantillerle gruplanan SGP hücreleri

Kantiller ile müşterileri gruplanmasında $5 \times 5 \times 5 = 125$ SGP hücreleri oluşmaktadır. Müşterileri segmente etmenin diğer bir yöntemi ise SGP bileşenlerini kümeleme yönteminde girdi olarak kullanmaktır.

10.1 SGP Bileşenlerinin Sürekli SGP Skoruna Dönüştürülmesi

SGP bileşenlerinin sürekli değerden oluşan bir skor olması, hem bu skorun yeni bir metrik olarak kullanımını ve kıyaslama kolaylığını sağlamaktadır. Bu mevcut skorların ağırlıklandırılması sonucu gerçekleşebilmektedir. Ağırlıklar bileşene verilen öneme göre farklılık gösterebilir. Dolayısı ile skor şu şekilde hesaplanmaktadır:

$$SGP\ skor = (S \times SW) + (G \times GW) + (P \times PW) \quad (10.1)$$

$SW = sonzamandardalık\ ağırlığı$

$GW = Sıklık\ ağırlığı$

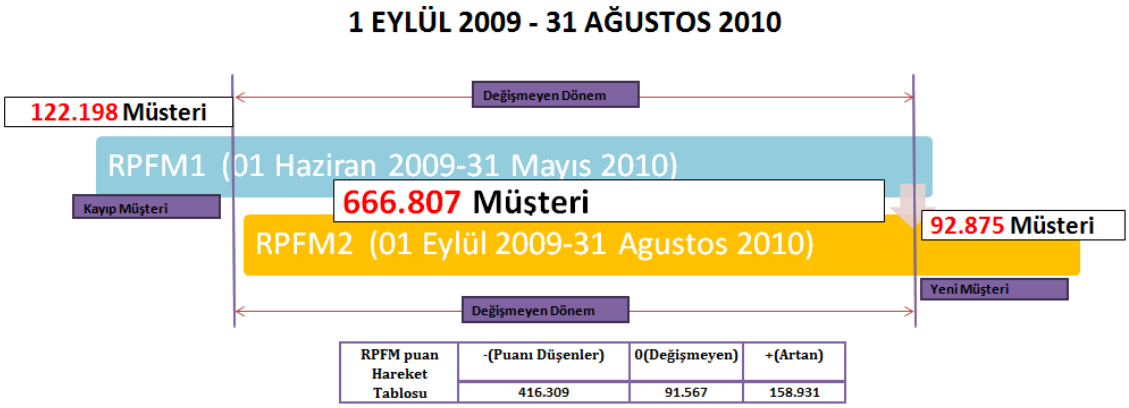
$PW = parasal\ ağırlığı$

Her bir bileşenin ağırlığı 10 olarak alınırsa, ölçek aralığı 30-150 arasında skorlar olacaktır. Bu skorlar da 0-1 arasında yeniden ölçümlenebilir:

$$yeni\ SGP\ skor = \frac{SGP\ skor - minimum\ SGP\ skor}{maksimum\ SGP\ skor - minimum\ SGP\ skor} \quad (10.2)$$

10.2 Gerçek Müşteri Verileri SGP Modeli

SGP değerlerini periyodik olarak hesaplanması ve değişimlerin takip edilmesi ve gözlemlenmesi her bir müşterinin işletmeye kattığı değerin takibi demektir. Diğer taraftan bu üç önemli ölçütün komple tek bir skorda birleştirilmesi müşterinin işletmeye göstermiş olduğu gerçek değerini yansıtabilmektedir. Periyodik takip ayrıca, müşteri skorlarında meydana gelecek azalmaları işletmeye uyarı olarak sunabilecektir. Bu durum işletmeye aksaklıkların hangi bileşenden kaynaklandığı bilgisini verecek ve işletmeye aksiyon alabilmek için tetikleme mekanizmaları oluşturulmasını sağlayacaktır. Sonuç olarak müşteri yaşam süresi değeri arttırılacak ve sadık müşteri süresi arttırılabilecektir. Ek olarak, parasal bileşeni müşterinin işletmeye sunduğu cirosal katkıyı ifade etmektedir. Ciro bazı durumlarda, özellikle yoğun kampanya, promosyon ve güçlü rekabet ortamlarında yapılan indirim vb. nedenlerden dolayı anlamlı bir değişken olamayabilmektedir. CRM uzmanı Dr. Hakan Tetik bu endişeyi ortadan kaldırabilmek adına **karlılık** etkisinin de bileşen olarak modele eklenmesi gerektiğini önermiştir. Şekil 10.3'de periyodik olarak (3' er aylık dönemlerde ötelenen(shift)) geçmiş 1 seneyi kapsayan müşteri işlem verilerinden (tekil müşteri tanımlayıcı numarayı içeren; alışverişlerini sadakat kart kullanan müşteriler) SGP modelinin(karlılık ile birlikte) sonuçları yer almaktadır. Şekil 9.3' den de görüleceği gibi 2. dönemde, 1. döneme göre skor değerleri düşen müşteri sayısı, 1. döneme göre skoru artan müşterilere göre oldukça fazladır. Bu nedenle bu müşterilerde meydana gelen skor düşüşlerinin sebepleri araştırılır; bu durum hangi bileşenden kaynaklanıyor? Eğer geliş sıklığı ise, geliş sıklığını arttırıcı bir teklif, eğer müşteri harcamasında bir düşüş varsa, alışveriş limitli teklifler sunmak, eğer müşteri sonzamandardalık bileşeninde sorun varsa elde tutma programları geliştirmek gerekecektir. Bunları gözlemlerken bileşenler arasında ilişki olabileceğine dikkat edilmelidir.



Şekil 10.3 Gerçek veriler ile SGP müşteri skor kıyaslaması

10.3 SGP Segmentasyon Prosedürü

Prosedür SGP bileşenlerinin hesaplanması ve müşterilerin SGP segmentlerine atanmasıdır. Uygulama adımları:

10.3.1 Veri Elde Edinim

İlgili döneme ait işlemsel veriler belirli veri temizleme süreçlerinden sonraki adıma hazırlanır. İşlemsel veriler işletmenin sadakat programına entegre olmalı ve her bir müşteri işlemi takip edilebilmelidir. Bu da sadakat programının doğru çalışması sonucu gerçekleşmektedir.

10.3.2 Segmente Edilecek Kitlenin Seçimi

Sadece aktif müşteriler SGP segmentasyonuna dahil edilmektedir (ilgili dönemde). İlgili zaman döneminde alışveriş yapmayan aktif olmayan müşteriler için elde tutma, geri kazanım, tekrar aktivasyon kampanyaları ile desteklenmelidir. Aynı zamanda yeni müşterilerde (ilgili döneme göre, belirli eşik zamandan önce gelen yeni müşteriler, örneğin 3 aydan önce gelen yeni müşteriler) güvenilir bir satın alma profili yansıtamayacakları sebebiyle analizden çıkarılabilir.

10.3.3 Veri Hazırlığı ve S, G ve P Ölçümlerinin Hesaplanması

Gruplanacak kitlenin seçimi yapıldıktan sonra, ilgili kitleye ilişkin SGP değerleri hesaplanır.

10.3.4 SGP Hücrelerinin Oluşumu, Kantil Gruplanma

SGP değerleri belirlendikten sonra sıralı değerleme yapılır ve hücreler belirlenir.

10.3.5 Kümeleme ile SGP Segmentlerinin Keşfedilmesi

Segmentasyon SGP değerleri hesaplandıktan sonra kümeleme yöntemleri uygulanabilmektedir. Elde edilen kümelerin profillerine göre küme isimleri belirlenip, başından beri amaç olan bire bir pazarlama anlayışını uygulayabilecek pazarlama aksiyonları ve müşteriyi anlamaya yönelik anlamlı bilgilere ulaşılmış olur. Kümeleme yerine SGP skorlarını süreklileştirerek değer bazlı segmentasyon gerçekleştirebiliriz, hatta bu segmentasyonu müşteri yaşam evreleri ile birleştirebiliriz.

Çizelge 10.1 "proc fastclus" prosedürü ile rastgele başlangıç prototipler ile gerçekleşen kümeleme

	Ortalama							
Değişken	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
N	**789.224**	**11**	**1.504**	**71**	**1**	**12**	**6**	**1**
DHT	599,8	96.725,90	15.410,10	37.490,70	265.279,70	61.586,10	131.567,30	184.042,60
DFNT	5,6	152	95,3	167,8	339	164,7	191,8	630
DINT	11,8	624,9	269,7	535,1	705	471,7	471	2.026,00
SFIL	240,9	56,5	33,1	33,2	101	103,3	43,7	0
BN	6,6	53,8	73,8	108	42	76,6	73,7	205
CN	2,8	5,8	7,4	7,7	5	6,6	6,7	8
DFF1	4,9	64,8	66,6	103,1	31	66,2	90,5	254
MN	3,1	11,2	11,1	11,3	8	8,7	10,3	12
SEG	7.526,20	109.389,70	148.098,40	230.641,50	64.821	131.233,20	112.720,50	380.238
UGN	5,3	22,4	38,4	48	12	32,4	32,7	67
UCN	6,8	48,7	69,9	95,3	34	67,8	60,5	161
VL	1,4	2,9	3,2	3,4	6	2,1	2,2	3
DC1HR	7%	22%	12%	15%	0%	17%	72%	6%
DC2HR	14%	47%	22%	22%	100%	36%	6%	19%
DCAHR	10%	2%	8%	7%	0%	10%	1%	5%
DCCHR	8%	1%	7%	7%	0%	5%	1%	7%
DCEHR	18%	10%	16%	15%	0%	12%	12%	34%
DCGHR	24%	16%	17%	20%	0%	6%	5%	19%
DCKHR	17%	2%	16%	14%	0%	4%	4%	9%

10.4 Gerçek Veriler- Örnek Olay İncelemesi

Çok katlı mağazacılık sektöründe hizmet veren bir ticari firmaya ait veriler ile farklı iki müşteri segmentasyon modeli kullanılarak elde edilen sonuçlar incelenmiştir.

10.4.1 K-Ortalamalar Kullanımı

Öncelikle Çizelge 10.1' de verilen datamart yapısı üzerinden bir segmentasyon modelini oluşturulmaya çalışılacak.

İlk önce müşterilerin doğal gruplarını keşfedebilmek için K-ortalamalar algoritmasının çalıştığı SAS üzerinde "Proc Fastclus" prosedürü ile kümeleme gerçekleştirilmiştir. K-ortalamaların en önemli problemi olan başlangıç kökleri (tohumlar) burada ortaya çıkmaktadır. İşletme baş edebileceği 8 küme oluşturmak istemektedir. Rastgele seçilen prototipler ile başlangıç kökleri belirlenmiştir.

*PHASE1;

PROC FASTCLUS DATA=FINDSEG.SEGDATA OUT= FINDSEG.SEGMENTS MAXCLUSTERS=7 MAXITER=0

/*SEED="RANDOM"*/OUTSEED= FINDSEG.TOH3 NOPRINT;VAR&VARMACROS;RUN;

Algoritma işlemi, küme karakteristiklerini yansıtan veriler ile şu sonuçlar çıkarmıştır:

Elde edilen kümelerdeki müşteri sayılarının oldukça dengesiz olduğu görülmektedir. Bu şekildeki bir küme yapısı içerisinde işletme herhangi bir aksiyon alamayacaktır. Üstelik bire bir pazarlama anlayışını yansıtmak isteyen ve bu bağlamda müşterilerini yönetebileceği anlamlı gruplara bölmek isteyen işletme, neredeyse müşterilerin tamamını 1. kümede toplamıştır. Algoritma pratikte kullanmak üzere anlamlı bir sonuç vermemiştir.

K-ortalamaların anlamlı küme yapıları ortaya çıkaramamasının en büyük sebebi başlangıç köklerinin belirlenmemesidir. Sadece k-ortalamalar değil, benzer şekilde hedef skor fonksiyon temelli kümelerde bu ve benzeri sorunlar olacaktır. Bu durum aynı zamanda segment profillerinin anlaşılmasında da sıkıntı teşkil etmektedir.

Diğer taraftan profil yapısı net olarak çıkarılamayan segmentler için bir segmentasyon yönetim strateji planı oluşturulması da mümkün olamamaktadır. Buna karşın hala k-ortalamalar algoritması ve türevlerinin işletme sorununa yanıt veremediği söylenemez.

Sorun başlangıç köklerinden kaynaklanıyorsa, çözümü başlangıç köklerini belirleyerek çözülebilir. Bu durum aslında işletmenin segment yapılarını kedisinin belirlemesi anlamına da gelmektedir. İşletme yönetmek istediği grup sayısında grup **prototiplerini belirleyerek** aksiyon alabileceği küme yapılarına sahip olabilmektedir. İşletmenin veya bir uzmanın kontrol edilebilecek ve pazarlama stratejilerine yön verebilecek küme yapıları oluşturması, doğal küme yapılarının oluşmasını beklemekten daha faydalı ve kolay bir çözüm yoldur. Sürekli değişmeyen sabit bir segmentasyon yapısı sağlar. Segmentasyonların kıyaslamasını ve müşteri segment geçişlerinin daha anlamlı takibi sağlanır.

Karmaşık veri yapılarına sahip ve birimleri bazında frekans düzeyi az olan perakende verilerinde uzaklık ölçülerinin müşteri farklılıklarını tamamen ortaya koyabileceği şüphelidir. Diğer bir sorun ise bu farklılıkların ne kadarını açıklayabildiğidir. Sahip olunan veriler arasında sayısal farklılıkların yanı sıra işletme bakış açısının ve kalitatif bilgilerinde bu bağlamda modele yansıması önemlidir.

Bu nedenle aslında K-ortalamaların dezavantaj gibi duran bu eksikliği aslında söylemek istediğimiz bu eksikliği belirli bir seviyede gidermede tamamlayıcı bir unsur olacaktır.

Sonuç olarak farklı bakış açılarının ortak bir noktada buluşacağı bir dizi başlangıç kökleri belirlenebilir. Bu kapsamda değerlendirilen proje, yeni kökler ile yeni bir kümeleme algoritması çalıştırılmıştır. Başlangıç kökleri fokus grup çeşitli Pazar araştırmaları ve uzman kişiler tarafından değerlendirilip girdi olarak k-ortalamalara yansımıştır.

Yeni kökler ile gerçekleşen yeni kümeleme sonuçlar Çizelge 10.2’ deki gibidir. Bu küme boyutları daha dengeli dağılmaktadır. Her bir kümenin kendi karakteristik özelliği bulunmaktadır. Seçilmiş değişkenlerde bu farklılıklar görülebilmektedir. Şekil 10.4’ de segmentlerin ortalama harcama, kişi sayısı ve toplam harcamalarına göre nasıl farklılık gösterdiği görülmektedir. Elde edilen kümeleme yöntemi, şimdi segmentasyon yönetim stratejisine uygun olabilmektedir. Artık model belirlendi. Şimdi veri tabanında müşteri segmentleri tanımlanmalı. Segmentlerin pazardaki yerlerinin belirlenmesi ve değerlendirilmesi. Her bir segmente öncelikli aksiyon için kar-zarar analizleri. Farklılaşmış stratejilerin kurulması. Kanal, servis, ürün, marka yönetimi.

PROC FASTCLUS DATA=FINDSEG. SEGDATAOUT= FINDSEG.SEGMENTS MAXCLUSTERS=8 MAXITER=0

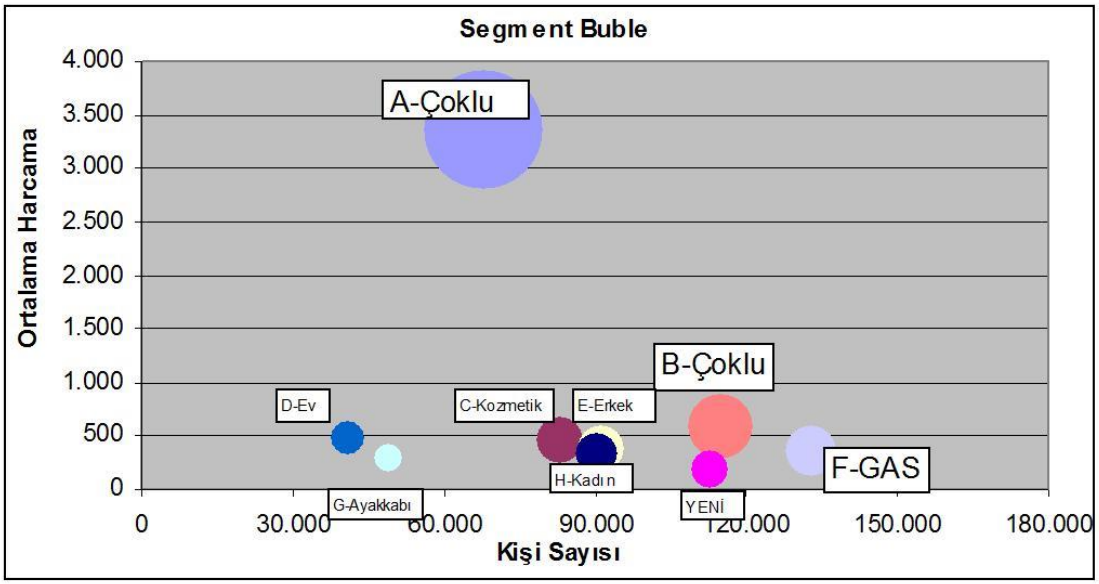
SEED=STATIC.TOH2_4SEG OUTSEED= FINDSEG.TOH2 NOPRINT;

VAR &VARIABLES;RUN;

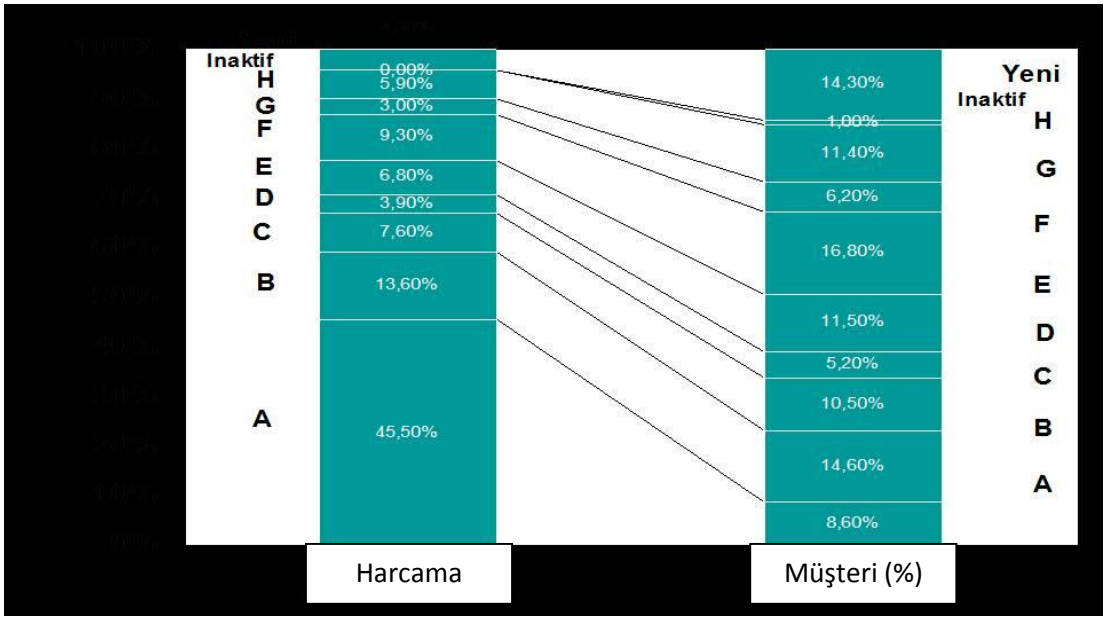
Çizelge 10.2 Subjektif bilgiler ile başlangıç prototiplerinin oluşturulması sonucu elde edilen küme değerleri

	Ortalama							
Değişken	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
N	67.729	115.083	83.078	40.804	91.163	132.970	49.006	90.208
DHT	3.372,23	593,14	460,19	478,8	373,8	353,1	306,48	331
DFNT	25,58	7,28	4,04	4,04	3,49	3,56	3,24	4,01
DINT	62,1	15,16	8,16	7,19	7,06	6,02	5,34	7,71
SFIL	89,06	223,7	276,45	307,49	315,48	298,34	318,01	296,22
BN	27,21	9,11	4,4	4,73	4,46	4,1	4,1	4,99
CN	6,19	3,9	2,38	2,75	2,18	2,19	2,49	2,43
DFF1	20,69	6,34	3,53	3,57	3,1	3,2	2,92	3,57
MN	8,45	4,28	2,75	2,74	2,44	2,51	2,36	2,71
SEG	41.357,39	10.099,59	5.451,20	5.426,62	4.101,91	3.755,28	4.077,03	4.926,24
UGN	18,58	7,75	3,73	4,29	3,55	3,47	3,52	4,13
UCN	27	9,36	5,89	5,22	4,28	4,02	3,99	5,03
VL	2,44	1,64	1,42	1,43	1,29	1,3	1,33	1,34
DC1HR	11%	5%	2%	**72%**	1%	2%	2%	2%
DC2HR	17%	6%	**79%**	4%	2%	2%	3%	4%
DCAHR	10%	8%	3%	3%	3%	3%	**71%**	4%
DCCHR	6%	32%	2%	3%	2%	2%	2%	3%
DCEHR	18%	15%	3%	5%	**80%**	6%	7%	5%
DCGHR	19%	14%	4%	6%	7%	**80%**	7%	5%
DCKHR	19%	14%	6%	6%	4%	4%	7%	**76%**

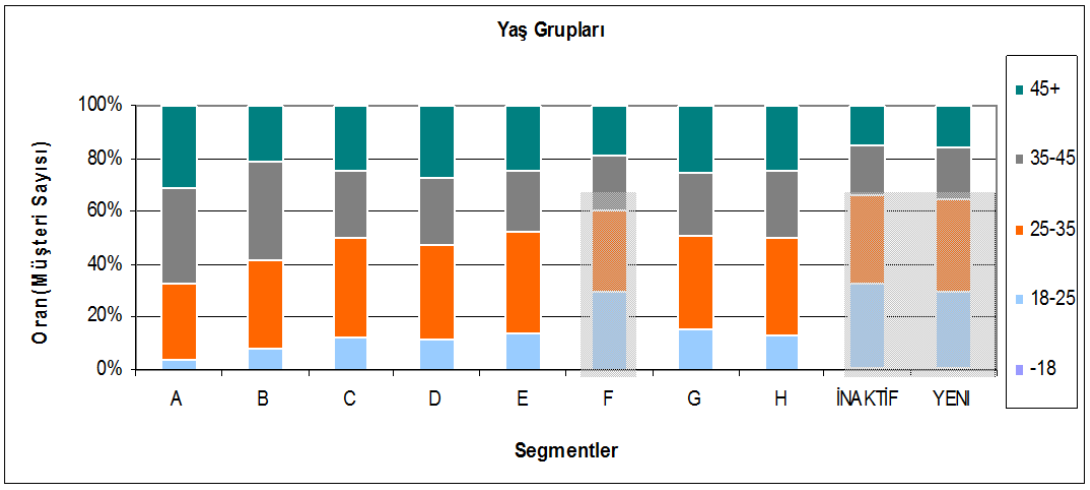
10.4.1.1 Profiller



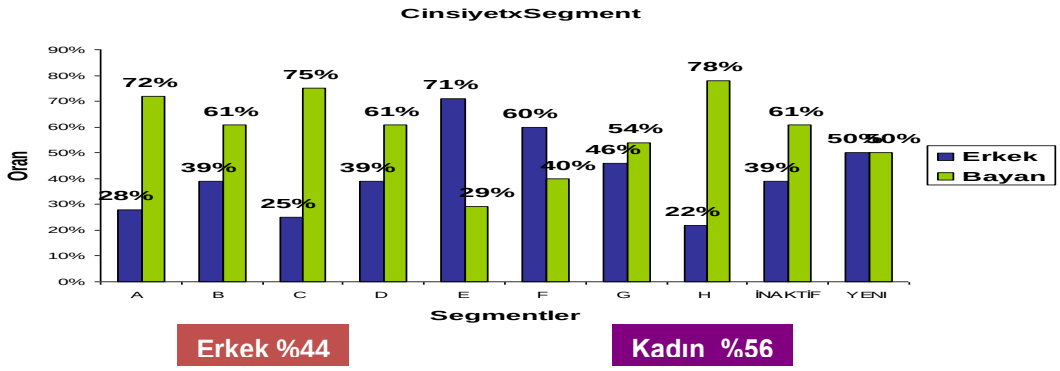
Şekil 10.4 Segmentlerin uzamsal gösterimi



Şekil 10.5 Subjektif prototipler ile seçilen segmentlerin ciro (%) ve müşteri sayısı (%)



Şekil 10.6 Segment yaş grup dağılımları



Şekil 10.7 Segment cinsiyet kırılımları

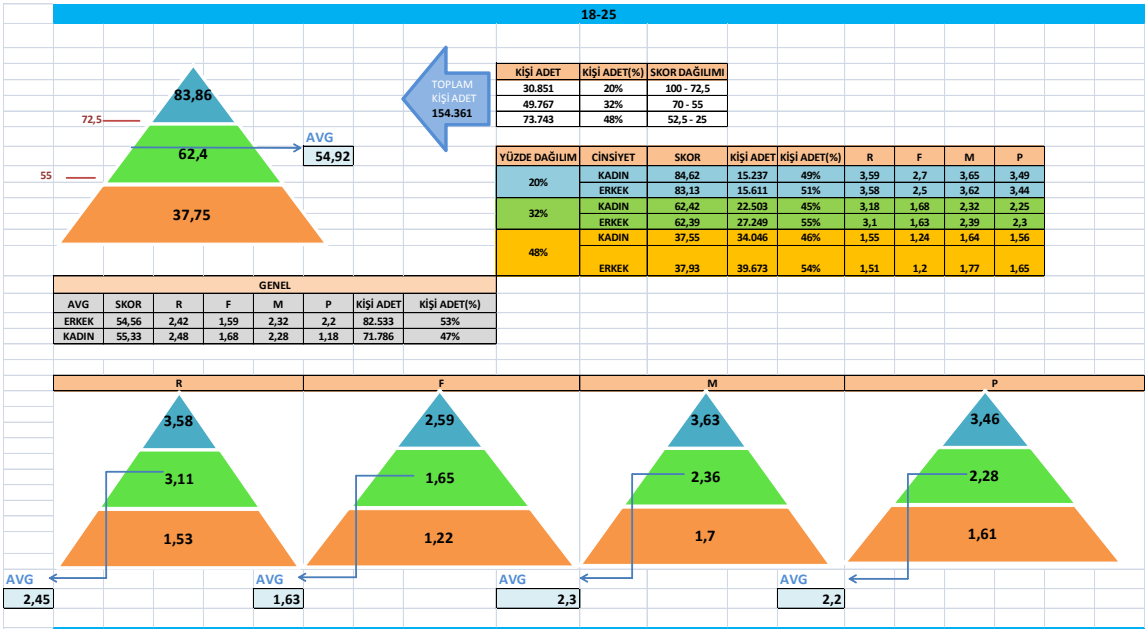
10.4.2 Yaşam Evrelerine Göre SGP Model Segmentasyon

Diğer bir yöntem ise SGP değer bazlı segmentasyonunun, yaşam evreleri ile birleştirilmesi sonucu oluşturulmuştur.“18-25”, “26-35”, “36-45” ve 46 yaş üzeri müşterilerin cinsiyet kırılımlarında SGP ve cirodaki handikap dolayısı ile karlılık bileşeni eklenerek SGPK skoru hesaplanmıştır. En yüksek %20’ lik skor “Altın”, sonraki %30’ luk skor “Gümüş” ve kalan %50’ lik kısım ise “Bronz” olmak üzere üç grup oluşturulmuştur. Sonuç olarak 4 yaş grubu, 2 cinsiyet kırılımı ve 3 değer kırılımı olmak üzere$4 \times 2 \times 3 = 24$ segment oluşmaktadır. Bu segmentler ayrıca diğer demografik bilgilerin (Meslek, hobi) toplanması ile mikro segment yapıları oluşacaktır. Bu mikro segmentler ile yaşam tarzları belirlenip, değer gruplarındaki değişimlere göre teklifler ve program kurguları geliştirilebilecektir.

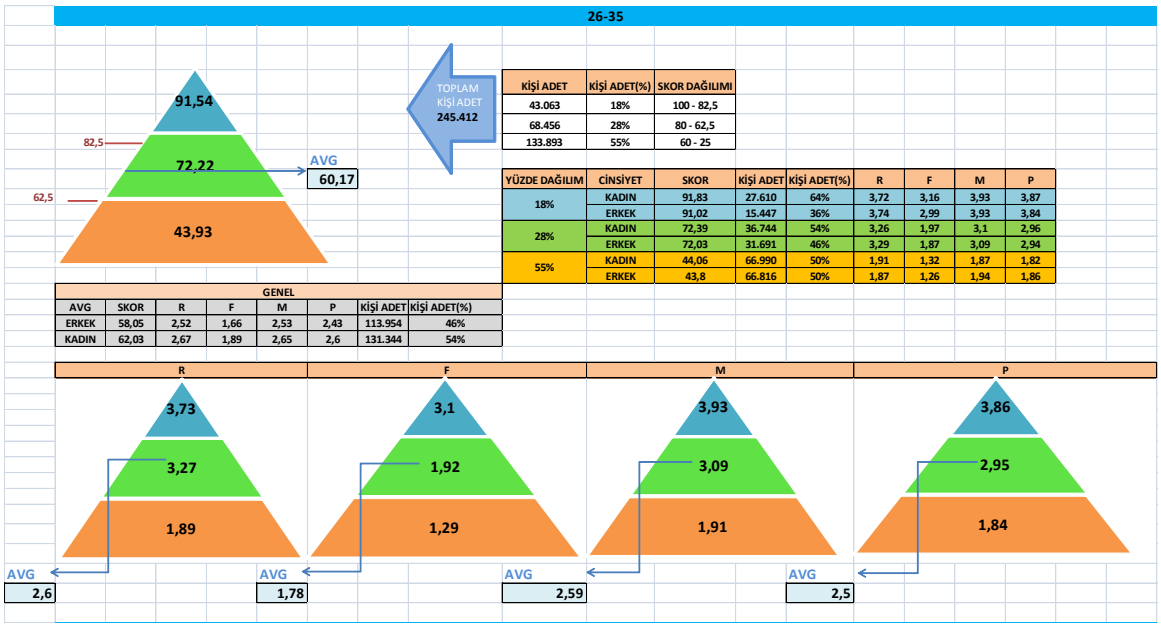
Model oldukça basit, yönetimi kolay, veri madenciliği algoritmaları içermeyen bir modeldir. Tamamen fokus grup çalışmaları, pazar araştırmalarından müşteri davranışları ve yaşam stilleri çıkartılarak derlenmiştir.

10.4.2.1 Profiller

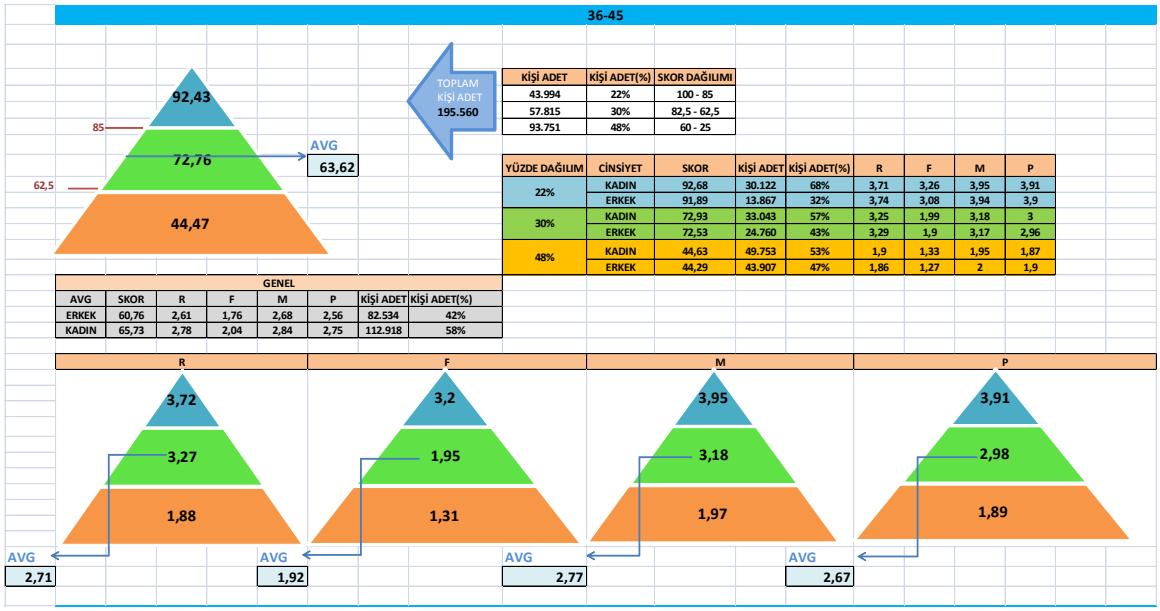
Şekil 10.8 “18-25” yaş grubu müşterilerinin skor değerlerine göre gruplandırılması



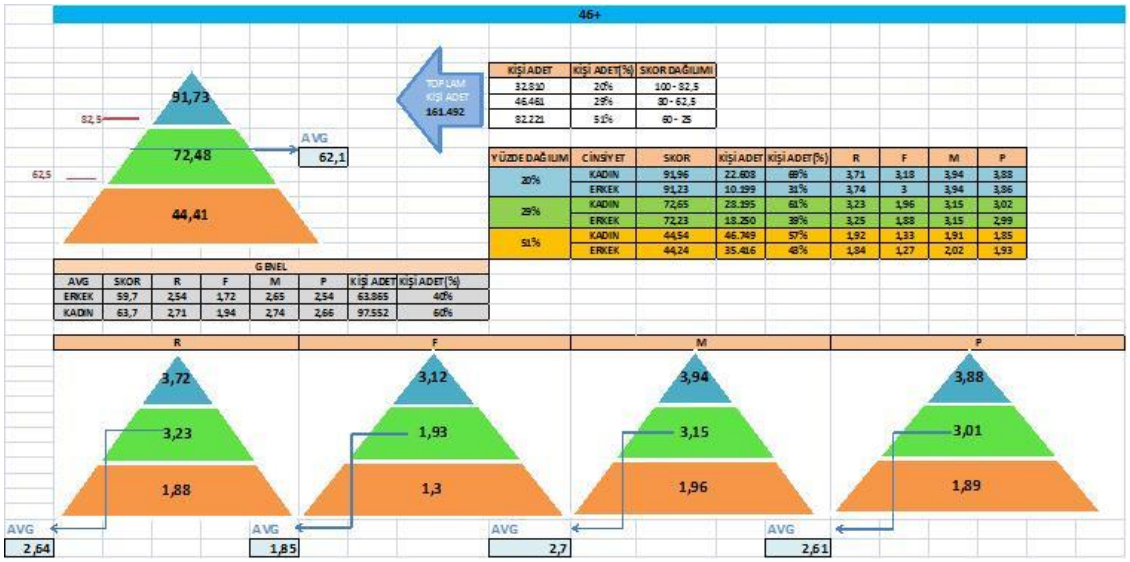
Şekil 10.9 “26-35” yaş grubu müşterilerinin skor değerlerine göre gruplandırılması



Şekil 10.10 “36-45” yaş grubu müşterilerinin skor değerlerine göre gruplandırılması



Şekil 10.11 “46 ve üzeri” yaş grubu müşterilerinin skor değerlerine göre gruplandırılması



BÖLÜM 11

SONUÇ VE ÖNERİLER

Veri madenciliği, perakendecilik sektöründe verilerin dijital ortama atılma kabiliyetlerinin gelişmesinden dolayı son zamanlarda oldukça yoğun bir şekilde kullanılmaktadır. Müşteri yaşam döngüsünde müşterinin nasıl elde tutulacağı, müşterinin ne kadar müşteri olma süresi olduğunu ve yeni müşteri edinimi gibi problemlerinde veri madenciliği yöntemleri çok değerlidir.

Firmalar müşteri ayrıştırma yöntemlerinde, farklı amaçlarda farklı müşteri kümeleme yöntemleri kullanmaktadır. SAS paket programı ile çalıştırmış olduğumuz fastclus prosedürü verilen müşteri hareket datasından, ilk prototiplerin rassal seçilmesinden dolayı; işletmeciye pazarlama aktivitesinde kullanması amacıyla anlamlı bir ayrıştırma sunamamıştır. Sonrasında prototip merkezlerinin pazarlama stratejisine ve firma gereksinimlerine uygun olarak derlenmesi ve prosedüre girdi olarak sunulması ile; işletmeci kendi ihtiyacına göre kullanabileceği anlamlı kümelere sahip olmuştur. Diğer taraftan uygulama tarafında oldukça basit olan RFM modellemesinde, müşteri skorları ile müşteriler belirli gruplara ayrıştırıldı. Bu yolla müşteriler, belirli yaş grupları ile ayrıştırılarak işletme için anlamlı sayıda segment kümeleri ve aksiyona dönüştürebilecek iletişim ve fırsat grupları oluşturuldu. Ayrıştırmanın fastclust prosedüründe, müşterileri davranış olarak modellemeye çalışılmıştır. RFM modeli ile de bu kitleler işletmeye olan değerleri bakımından modellenmiştir.

Sonuç olarak deneyimsel bir girdi sunarak ayrıştırma işlemi uygulama açısından anlamlı hale getirilmiştir. Bu bilgilerle hem müşteri davranış ve ihtiyaçları belirlenmiş, hem de işletmenin tüm müşteri kitlesine bir aksiyon almak yerine, belirli segment grupları ile,

daha hedef odaklı bir pazarlama anlayışına sahip olması sağlanmıştır. Müşteri ayrıştırma firmaya iletişiminde farklılaşmasını ve müşteriye gerçekten ihtiyacı olan ürünü sunması için fırsat oluşturmaktadır.

Uygulamanın gerçekleştirdiği veriler Türkiye' nin çok katlı mağazacılık sektöründe önde gelen firmalarından biridir. Giyimden kozmetiğe, ayakkabıdan çocuk ürünlerine yaklaşık 1000' e yakın markanın barındırıldığı 7 farklı kategoride Türkiye' de perakendecilik sektöründe yeralan bir firmadır. Firma gerçekleştirmiş olduğu sadakat programı ile kişiye özel fırsatlar ve öncelikler sunmaktadır. Kişiselleştirilmiş sadakat kartı ile müşterileri bu fırsatları sunmaktadır.

Uygulamada, davranış bazlı modelleme sonrasında, 9 farklı segment oluşmuştur;

Çoklu Kategori Segmenti 1 (67.729 müşteri): Ortalama 5 kategoriden de alışveriş yapmış müşterilerdir. Çoklu kategori alışverişi ile ortalama harcama tutarlarıda yüksek seviyededir. Bu segment bu nedenlerle, diğer segmentlerden anlamlı derecede ayrışmaktadır. Müşteri bilgileri bu segmentin ağırlıklı 35 yaş üstü müşterilerden oluştuğunu ve müşteri oranı %8,6 olmasına karşın, harcama oranı %45,5' dir. Ortalama ayda iki kez alışveriş gerçekleştirmektedirler.

Çoklu Kategori Segmenti 2 (115.083 müşteri): Çoklu segment 1' e en yakın özellikleri gösteren segmenttir. 1' e göre, harcama potansiyeli daha az olup, yine ortalama 4-5 farklı kategori harcaması yapmaktadırlar.

Kozmetik Segmenti (83.078): Harcama oranına bakıldığında alışverişlerinin %80 gibi bir kısmını müşteriler kozmetik reyonlarından gerçekleşmektedir. Ağırlıklı olarak bayan müşterilerdir.

Erkek Segmenti (91.163): Harcamalarının yüksek bir oranını Erkek reyonlarından gerçekleştiren müşterilerdir. Özellikle klasik giyimcilerin yeraldığı bir segmenttir.

Kadın Segmenti (90.208): Firmayı Kadın reyonlarından dolayı tercih eden müşteriler olup, sıklıkla ziyaret eden müşterilerdir.

Genç Aktif Spor Segmenti (132.970): Müşteri hacmi olarak en çok kitle bu segmenttedir. Ağırlıklı olarak 35 ve altı yaş grubudur.

Bu şekilde oluşturulan müşteri kitleleri ile, milyonlarca kişiye milyonlarca iletişim ve aksiyon almak yerine, bu 9 kümede yer alan kitlelere, 9 iletişim ve aksiyon alarak; maliyetin minimize edilmesi, zaman kazancını da maksimize edilmesi sağlanmıştır.

Diğer taraftan, veriler satış hareket verileri ve sadakat kartı ile birleştirilen verilerden oluşması sebebiyle; sadakat kartı ile gerçekleşmeyen satışların analiz dışı olduğu anlaşılmaktadır. Sadakat kartı ile gelen müşteri bilgiler (bu bilgiler müşterinin en az bir iletişim bilgisine sahip olunan bilgilerdir.) özellikle analiz sonrasında, pazarlama aksiyonlarında daha değerli olmaktadır. Bu noktada kritik konu, müşteri herhangi bir alışverişinde kartını kullanmadığında, bu alışverişi analize girememektedir. Ayrıca her müşterinin bir sadakat kartı olmayabilmektedir. Bu nedenle müşteri alışveriş sırasında mutlaka sadakat kartını kullanmalı veya daha pratik yöntemler ile müşteri alışverişinde tanınmalıdır. Bütün bu handikaplara karşı uygulama verilerinde ilgili dönemde kart kullanım penetrasyonu %53 oranındadır. Dolayısı ile firmanın müşterilerini anlamlı bir oranda tanıyabildiği söylenebilir.

Perakende dinamiklerinde pazarın giderek daralması, rekabetin artması müşteriyi elde tutmak giderek zorlaşmaktadır. Bu nedenle şirketler hangi müşterilerinin risk grubunda olduğunu merak etmektedirler. Telekomünikasyon, Call Center ve medikal sektör dışında, perakende sektöründe yeterli veri imkanı olmasa dahi; churn modelleri geliştirmektedir. Firmalar pazarlama aksiyonlarını, kaybetmek üzere olduğu veya kaybettiği müşterileri üzerinde de almaktadır. Gerçekleştirmiş olduğumuz çalışmada, hangi segment müşterilerin riskli olduğuna dair çıkarsamalarda bulunulmuştur.

Ortalama olarak, 6.4 aydır gelmeyen müşteriler pasif müşteriler olarak belirlenmiştir. Bundan sonraki aşamada, pazarlama ve CRM tarafında bu müşterilere aksiyon belirlenmesi gerekmektedir.

Elde edilen ayrıştırma yöntemleri ile gerçekleştirilecek aksiyonlar ile firma benzer özellikler gösteren müşterilerine, aynı mesajlar ile erişebilecektir. Bundan sonraki çalışmalarda, bu tez içerisinde yer verilmeyen denetimli (supervised learning) modellerin perakende firmalarındaki uygulamalarının yeralması gerektiği görüşündeyiz. Diğer taraftan tüketicilerin yaşamlarına daha yoğun bir şekilde giren mobil aletlerdeki gelişmeler, daha dinamik pazarlama anlayışı getirmektedir. Bu durum veri madenciliğinin de daha dinamik yapıda olmasını gerektirmektedir.

KAYNAKLAR

[1] Girolami, M., (2002). "Mercer Kernel-Based Clustering in Feature Space", IEEE Transactions On Neural Networks, 13(3):780-784.

[2] Giudici, P. ve Figini, S., (2009). Applied Data Mining for Business and Industry, Second Edition, John Wiley & Sons, Chichester.

[3] Greenlaw, R. ve Kantabutra, S., (2010). Introduction to Clustering: Algorithms and Applications, IGI Global, MA.

[4] Hand, D., Mannila, H. ve Smyth, P., (2001). Principles of Data Mining, A Bradford Book The MIT Press, London.

[5] IBM redbooks, (2001). Mining Your Own Business in Retail Using DB2 Intelligent Miner for Data, First Edition, IBM International Technical Support Organization, California.

[6] Cios, K. J. , Pedrycz, W., Swiniarski, R. W. ve Kurgan, L., (2007). Data Mining A Knowledge Discovery Approach, Springer Science + Business Media LLC, New York.

[7] Berry, M., Linoff, S. G., (2004). Data Mining Techniques For Marketing, Sales, and CRM, Second Edition, Wiley Publishig Inc., Indiana.

[8] Han, J. ve Kamber, M., (2006). Data Mining: Concepts and Techniques, Second Edition, Elsevier's Science & Technology Rights Department, Oxford.

[9] Olson, D. ve Delen, D., (2008). Advanced Data Mining Techniques, Springer, Berlin.

[10] Maimon O. and Rokach, L., (2010). Data Mining and Knowledge Discovery Handbook, Springer, London.

[11] Pyle, D., (1999). Data Preparation for Data Mining, Morgan Kaufmann Publishers Inc, San Francisco.

[12] Frank, R. E., (1972). Market Segmentation, Prentice-Hall, N.J.

[13] Tsiptsis, K. and Chorianopoulos, A., (2009). Data Mining Techniques in CRM: Insight Customer Segmentation, John Wiley & Sons Ltd, London.

[14] Witten, I.H., Frank, E. and Hall, M.A., (2005). Data Mining Practical Machine Learning Tools and Techniques, Elsevier Inc., San Francisco.

[15] Kantardzic, M., (2011). Data Mining: Concepts, Models, Methods, and Algorithms, Second Edition, John Wiley & Sons, New York.

[16] Bozdogan, H., (2004). Statistical Data Mining and Knowledge Discovery, A CRC Press Company, Washington.

[17] Allison, P. D., (2010). Survival Analysis Using SAS: A Practical Guide, Second Edition, SAS Institute, North Carolina.

[18] Awad, M., Khan, L., Thuraisingham, B. and Wang, L., (2009). Design And Implementation Of Data Mining Tools, CRC Press, New York.

[19] Halgamuge, S.K. and Wang, L., (2005). Classification and Clustering for Knowledge Discovery, Springer, Berlin.

[20] IBM Redbooks, (2002). Enhance Your Business Applications: Simple Integration of Advanced Data Mining Functions, First Edition, IBM International Technical Support Organization, New York.

[21] Nemati, H. and Barko, C. D., (2004). Organizational Data Mining: Leveraging Enterprise Data Resources for Optimal Performance, Idea Group Publishing, Pennsylvania.

[22] Mitsa, T., (2010). Temporal Data Mining, Taylor and Francis Group, London.

[23] Linoff, G. S., (2004), Survival Data Mining for Customer Insight Data Miners, http://www.data-miners.com, 1 Mart 2010.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı	: Gürkan MOCAN
Doğum Tarihi ve Yeri	: 01.03.1981 - Sivas
Yabancı Dili	: İngilizce
E-posta	: gurkanmocan@gmail.com

ÖĞRENİM DURUMU

Derece	**Alan**	**Okul/Üniversite**	**Mezuniyet Yılı**
Lisans	İstatistik	Yıldız Teknik Üniversitesi	2004
Lise	Fen	Bahçelievler Lisesi	1998

İŞ TECRÜBESİ

Yıl	**Firma/Kurum**	**Görevi**
2008-2010	Boyner Büyük Mağazacılık	CRM Yöneticisi
2006-2007	ARC	Proje Yöneticisi
2005-2006	BG Marketing	Proje Yöneticisi