

VERİ MADENCİLİĞİNDE KARAR AĞAÇLARI VE BİR SATIŞ ANALİZİ UYGULAMASI

Gül Gökay EMEL

Uludağ Üniversitesi, İİBF, Bursa

Çağatan TAŞKIN

Uludağ Üniversitesi, İİBF, Bursa

ÖZET

Günümüzün gelişmiş bilgisayar teknolojisi işletmelerde veri yığınları oluşturmaktadır. Stratejik pazarlama kararlarının doğru ve zamanında alınabilmesi, bu verilerin en iyi şekilde analiz edilebilmesi ve anlamlı hale getirilmeleri ile mümkündür. Bu nedenle, bunu sağlayacak veri madenciliği gibi güçlü araçlara ihtiyaç vardır. Veri madenciliği ile tahmin modellerinin oluşturulmasında en yaygın ve etkin olarak kullanılan yöntemlerden biri karar ağaçlarıdır. Bu çalışmada, ele alınan bir perakendeci işletme için işletmenin müşterilere göre kişiselleştirilmiş satış hareketlerini içeren veri tabanından yararlanarak ayrıntılı ve göreceli ölçüm sonuçlarını içeren bir satış analizi amaçlanmaktadır. Oluşturulacak sınıflama tipi satış tahmin modeli için C&RT karar ağacı tekniğinden yararlanılmaktadır. Söz konusu modelden ayrıntılı satış değerleri, müşterilerin satın alma davranışı profilleri ile birlikte elde edilebileceğinden işletmenin pazarlama stratejilerinin etkinliği de arttırılabilecektir.

Anahtar Kelimeler : Veri madenciliği, Karar ağaçları, Satış Analizi

DECISION TREES IN DATA MINING AND A SALES ANALYSIS APPLICATION

ABSTRACT

Today's highly developed computer technology helps to accumulate large amounts of data in enterprises. Making appropriate marketing decisions on time can be succeeded with the help of the analysis of these data. That is why, robust tools such as data mining are needed. Decision trees is one of the data mining methods which is used widely and efficiently in prediction. In this paper, a sales analysis that includes the customers' sales transactions of an enterprise in detail, is aimed. C&RT decision tree technique is used for the sales prediction

model. The results of the model are important from the marketing perspective because they will help to increase the efficiency of the strategies.

Keywords: Data mining, Decision trees, Sales analysis

1.GİRİŞ

Verilerin analizinde istatistik, pek çok problemi çözmeye önemli bir araçtır. Ancak, bazı durumlarda kullanımı sınırlıdır. Bu durumlarda, yapay sinir ağları, kural çıkarımı, mantık programlama, karar ağaçları, genetik algoritma gibi akıllı veri analiz yöntemlerini içeren veri madenciliğine gereksinim ortaya çıkmaktadır (Ryu ve Eick, 2004). Ancak, bu yöntemlerden bazıları kara kutu yaklaşımları şeklindedir. Başka bir ifade ile açıklama getirmeden iyi tahmin yaparlar. Bu da, bu yöntemlerin zayıf yönünü oluşturur. Karar ağacı yaklaşımı ise bu zayıflığın söz konusu olmadığı veri madenciliği yöntemlerinden biridir (Zorman vd., 2001, 109-110).

Karar ağacı yaklaşımı, hedef fonksiyonlarını yaklaşık olarak hesaplamak için kullanılan ve öğrenme fonksiyonunun karar ağacı ile gösterildiği bir yöntemdir. Bir karar ağacı ise ağaç görünümünde tanımlayıcı ve tahmin edici bir modeldir (Ma, 1998, 44). Bu model, karar alıcıya karar alırken hangi faktörlerin göz önüne alınması ve her bir faktörün kararın farklı çıktıları ile geçmişte nasıl ilişkili olduğunun belirlenmesi konularında yardımcı olur (Bounsaythip ve Esa,2001,20).

Karar ağaçları, veri madenciliğinde kurulumlarının ucuz olması, yorumlanmalarının kolay olması, veri tabanı sistemleri ile kolayca bütünleştirilebilmeleri ve güvenilirliklerinin tatmin edici düzeyde olması gibi nedenlerden veri madenciliğinin sınıflama modelleri arasında yaygın kullanıma sahiptir. Karar ağaçlarının işletme problemlerindeki iyi bilinen uygulamalarına, mektupla yapılan pazarlamada hangi demografik grupların yüksek cevaplama oranına sahip olduğunun belirlenmesi, bireylerin kredi geçmişlerini kullanarak kredi kararlarının verilmesi, geçmişte işletmeye en faydalı olan bireylerin özelliklerini kullanarak işe alma süreçlerinin belirlenmesi, hangi değişkenlerin satışları etkilediğinin ve hangi değişkenlerin ürün hatalarına yol açtığı belirlenmesi gibi bir çok örnek verilebilir.

Bu çalışmada, işletmelerin satış hedeflerine ilişkin fiili satışlarını ölçme ve pazardaki başarılarını değerlendirme çabaları olan satış analizine farklı bir yaklaşım ele alınmaktadır. Satış analizi için satış hacmi analizi ve pazar payı analizi önemli iki araçtır. Makro satış hacmi analizi ayrıntılı değildir, mutlak ölçüm yapar ve genellikle işletmelerin yönetim bilgi sistemlerinden rutin rapor şeklinde elde edilir. Bu nedenle günümüzde yetersiz kalmaktadır. İlişki pazarlaması ve analitik CRM çalışmalarıyla ise çok çeşitli ve ayrıntılı bilgi elde edilebilmektedir. Ancak, bu çalışmalar kapsamlı veri organizasyonları ve iyi düzenlenmiş müşteri veri tabanı gerektirmektedir. İyi bir müşteri veri tabanı ise satış verileri ile bütünleştirilmiş, müşteri karakteristiklerini, kısaca onların kim olduklarını, işletme kararlarına

cevaplarını (ne, ne zaman, nerede satın alınmış) ve satın alma tarihçesi şeklindeki üç temel veri grubunu kapsar.

Anlamli mikro satış hacmi analizleri için de bu tip veri tabanlarının içerdiği verilere ihtiyaç vardır. Bu veri tabanları ile veri madenciliği yöntemlerinden yararlanarak işletme için yüksek kırılımlara sahip ve gerekli göreceli ölçümlerin olduğu satış analizleri yapılabilir. Bu çalışmada, bir işletmenin mevcut müşterilerinin belli bir dönemdeki toplam harcama tutarları, satın alınan ürünlere göre C&RT karar ağacı tekniği ile sınıflamaya tabi tutularak bir mikro satış hacmi analizi yapılmaktadır. Amaç, işletmenin toplam satış değerini müşterilerin satın alma örüntülerine dayalı olarak ayrıntılı ve göreceli olarak analiz etmektir.

2.VERİ MADENCİLİĞİ VE TEKNİKLERİ

Veri madenciliği, büyük miktardaki gözlenmiş verilerden kuralların, örüntülerin ve modellerin ortaya çıkarılmasıdır (Shaw vd., 2001, 128). Bir başka ifade ile veri madenciliği, veri tabanları veya veri ambarlarında yer alan yığın veri içindeki gizli örüntüleri ve ilişkileri bulmak için istatistiksel algoritmaları ve yapay zeka yöntemlerini kullanan karmaşık bir veri arama yeteneği olarak tanımlanabilir (Gargano ve Raggad, 1999, 81-82). Veri madenciliği; aynı zamanda bilgisayar bilimini, makine öğrenmesini, veritabanı yönetimini, matematiksel algoritmaları ve istatistiği birleştiren disiplinler arası bir alandır (Liao, 2003, 157).

Veri madenciliğinde örüntü tanıma faaliyetleri üç temel sınıfta toplanabilir. Bunlar; keşif (discovery), tahmin edici modelleme (predictive modelling) ve adli analizdir (forensic analysis). Keşif, bir veri yığınınındaki gizil örüntüleri önceden belirlenmiş bir fikir veya hipotez olmadan ortaya çıkarma sürecidir. Başka bir ifade ile verilerin içinde saklı olarak bulunan, hangi ürünlerin birlikte satıldığı veya hangi grup müşterilerin hangi zaman aralıklarında bir hizmeti kullandıkları gibi davranışları ortaya çıkarmaya yarar. Tahmin edici modelleme, ortaya çıkardığı örüntüler ile geleceği tahmin etmede kullanılmaktadır. Başarılı bir kredi verme işlemi veya bir hata olasılığı belirleme işlemi tahmin edici modelleme ile gerçekleştirilebilmektedir. Adli analiz ise ortaya çıkarılmış örüntülerin, kural dışı veya anormal veri elemanlarını bulmak için kullanılması süreci olarak tanımlanabilir (Rygielski vd., 2002; 487-488; Yun vd., 2003; 182).

Veri madenciliğinde tahmin edici modeller ile örüntü tanıma işi sınıflama, regresyon ve zaman serileri yaklaşımlarını içerir. Bu modeller, neyin tahmin edilmesinin istendiğine dayalı olarak farklılaşırlar. Çıktı niteliğinin sürekli değerleri için tahmin istenir ise regresyon analizi, zamanın ayırt edici özellikleri ile ilgileniliyor ise zaman serileri, iyi veya kötü gibi az sayıdaki ayrık kategoriye sahip bir özel veri ögesi için bir kategorikal değer tahmini yapılmak isteniyor ise sınıflama gerekir. Eldeki verinin gruplarını bulan kümeleme, birliktelik ve ardışıklık kurallarını elde etmeyi kapsayan birliktelik analizi ve ardışıklık keşfi davranışı ise tanımlama amaçlı kullanılır (Rygielski vd., 2002, 488; Moshkovich vd., 2002, 303-304).

Veri madenciliği tekniklerinin ayrıntılı bir sınıflaması verinin saklanması temel alınarak da yapılabilir. Bir başka ifade ile veri madenciliği teknikleri, eldeki verinin madencilik süreci uygulandıktan sonra saklanıp saklanmamasına göre bir sınıflamaya tabi tutulabilir. En yakın komşu ve duruma dayalı usullama (case-based reasoning) gibi veri madenciliği tekniklerinde veri kümesi gelecekteki bir örüntü eşleştirmesi için saklanmaktadır. Saklama tabanlı teknikler, yalnızca tahmin edici modelleme ile adli analiz görevlerini uygularlar ve ortaya çıkarılan örüntüleri süzmediği için özbilgi keşfi yapmazlar. Örüntü süzme tabanlı teknikler ise üç sınıfa ayrılabilir. Bunlar; mantıksal, çapraz tablolama ve denklemsel sınıflar olarak sıralanabilir. Bu teknikler, bir veri kümesi içinde mevcut olan çeşitli örüntüleri ortaya çıkarır ve daha sonra bu örüntüler çeşitli amaçlar için kullanılır. Bu teknikler, ayrıca ortaya çıkarılabilecek örüntü tipleri ve bunların tanımlanmaları ile ilgili sorulara yanıt arar. Mantıksal sınıfa ait teknikler hem sayısal hem de sayısal olmayan verileri kullanabilirken, denklemsel sınıfa ait teknikler yalnızca sayısal veriler üzerinde işlem yapabilir. Çapraz tablolama sınıfına ait teknikler ise yalnızca sayısal olmayan veriler üzerinde çalışabilir (Rygielski vd., 2002, 490-491; Bounsaythip ve Esa, 2001, 9-11).

3. SINIFLAMA VE KARAR AĞAÇLARI

3.1. Sınıflama

Sınıflama, veri madenciliğinin en tanınmış işidir. Girdilerin çeşitli niteliklere göre bir sınıflayıcı (model) tarafından sınıflara atanması sürecidir. Eldeki nesnelere bir sınıfa atanıp atanmayacağına ya da sınıflardan hangisine atanacağına belirlenmesidir. Başka bir ifade ile nesnelere veya durumlar için uygun sınıf tahmin edilmesidir. Sınıflama girdileri, her biri bir sınıf etiketi ile etiketlenecek gözlem veya örneklerden oluşan bir eğitim kümesidir. Çıktı ise modelin her bir gözleme niteliklere dayalı olarak atadığı sınıf etiketidir.

Sınıflama, makine öğrenmesinin de önemli araçlarından biridir. Tümevarımsal öğrenmenin temel çerçevesi eğitim örneklerinden oluşan bir eğitim kümesi ile test örneklerinden oluşan bir test kümesini içerir. Sınıflama iki adımda gerçekleşir. Bunlar verilerin eğitimi ve modelin testidir. Eğitim, eğitim kümesinden çıkarımla modelin oluşturulması, test ise test kümesini kullanarak modelin kesinliğinin kontrol edilmesidir. Modellerin kesinliğinin belirlenmesi için test örneklerinin iyi bilinen sınıfı, model tarafından tahmin edilen sınıf ile karşılaştırılır. Test örneklerinin model tarafından doğru olarak sınıflanma oranı kesinlik oranını verir. Girdilerden bu çıktıları üreten model, daha sonra sınıf etiketi bilinmeyen veya kayıp olan yeni gözlem veya örneklerin sınıf etiketini tahmin etmek için kullanılabilir. Ana sınıflama teknikleri olarak lojistik regresyon, diskriminant analizi, karar ağaçları, duruma dayalı usullama, Bayesgil sınıflayıcıları, eğer-sonra kuralları (kural çıkarımı), diğer mantıksal formüller yapay sinir ağları, bulanık kümeler, kaba kümeler sayılabilir.

Sınıflama ile ilgili olarak literatürde, farklı bir yaklaşım da söz konusudur. Bu yaklaşıma göre sınıflama; tam sınıflama ve kısmi sınıflama olmak üzere ikiye

ayrılmaktadır. Buradaki tam sınıflama kavramı veri içindeki tüm sınıflar ve örnekleri kapsayan modeller ile ilgilidir. Bunlara örnek olarak, yapay sinir ağları, C&RT, CHAID, C4.5, C5.0 ve diğer karar ağaçları gibi akıllı teknikler ile diskriminant analizi gibi istatistiksel araçlar verilebilir. Kısmi sınıflamada, tam sınıflamada olduğu gibi veri sınıflarının özellikleri gösterilmektedir. Ancak kısmi sınıflandırma modellerinde tüm sınıflar veya verilen sınıfın tüm örnekleri kapsamayabilir. Bu modellerden biri olarak birliktelik kuralları verilebilir (Bloemer vd., 2003; 119-120).

3.2. Karar Ağaçları ve Ağaç Tümevarımı

En önemli sınıflama araçlarından biri olan karar ağaçlarında, öğrenme algoritması basittir. Ortaya konan özbilginin gösterimi kolaylıkla anlaşılabilir. Karar ağaçları yalnızca kararları göstermezler, aynı zamanda kararların açıklamasını da içerirler. Karar ağacını oluşturan eğitim süreci tümevarımdır. Bir eğitim nesnelere kümesinden karar ağacı oluşturma yordamı, ağaç tümevarımı (tree induction) olarak adlandırılır. Ağaç tümevarımı yöntemi öz bilgi keşfinin en yaygın yöntemlerinden biridir. Sınıflama veya tahmin için kullanılacak ağaç benzeri örüntüleri keşfetme için bir yöntemdir.

Karar ağaçları bilgi keşfi sırasında pek çok test gerçekleştirerek, hedefi tahmin etmede en iyi sırayı bulmaya çalışırlar. Her bir test karar ağacındaki dalları oluşturur ve bu dallar da diğer testlerin gerçekleşmesine neden olur. Bu durum, test işleminin bir yaprak düğümünde (leaf node) sonlanmasına kadar devam eder. Kökten hedef yaprağa kadar olan yol, hedefi sınıflandıran “kural” olarak adlandırılır. Kurallar “eğer-sonra” (if-then) yapısındadır (Bounsaythip ve Esa, 2001, 18-19). Karar ağacına dayalı analizler (www.spss.com/classification_trees);

- Belirli bir sınıfın olası üyesi olacak elemanların belirlenmesinde (segmentation),
- Çeşitli vakaların yüksek, orta, düşük risk grupları gibi çeşitli kategorilere ayrılmasında (stratification),
- Gelecekteki olayların tahmin edilebilmesi için kurallar oluşturulmasında,
- Parametrik modellerin kurulmasında kullanılmak üzere çok miktardaki değişken ve veri kümesinden faydalı olacakların seçilmesinde,
- Sadece belirli alt gruplara özgü olan ilişkilerin tanımlanmasında,
- Kategorilerin birleştirilmesinde ve sürekli değişkenlerin kesikli değişkene dönüştürülmesinde

yaygın olarak kullanılmaktadır.

3.3. Karar Ağacı Tümevarımının Adımları

Bir karar ağacı tümevarımına, boş bir ağaç ve eğitim kümesinin tamamı ile başlanır. Tümevarım yordamı yinelemeli bir süreçtir ve her bir yineleme dört adımdan oluşur. Bu adımlar aşağıdaki gibi açıklanabilir (Zorman vd., 2001,110):

- Adım 1:** Eğer eğitim kümesindeki tüm eğitim nesnelere aynı sonuca sahip ise bu sonuç ile bir yaprak oluşturulur ve dördüncü adıma geçilir.
- Adım 2:** Bulgusal değerlendirme fonksiyonunun yardımı ile kökten mevcut düğüme kadar olan yol üzerinde henüz kullanılmamış tüm nitelikler arasından en iyi nitelik bulunur. Seçilen nitelik üzerinde bölünme işlemi ile içsel bir düğüm yaratılır. Sonra eğitim kümesi alt kümelere ayrılır.
- Adım 3:** Eğitim nesnelere her bir alt kümesi için birinci adıma gidilir.
- Adım 4:** Bir düzey yukarı çıkılarak adımlar tekrar edilir.

Yukarıdaki adımlardan da anlaşılacağı gibi karar ağacı algoritmalarında iki temel işlem gerçekleştirilmektedir. Bu işlemler; bölme (splitting) ve budama (pruning) işlemleridir. Algoritmaların sonlanması ise uygulanan durdurma kriterine göre olmaktadır. Bunlar kısaca şöyle açıklanabilir (Bounsaythip ve Esa, 2001, 18-19):

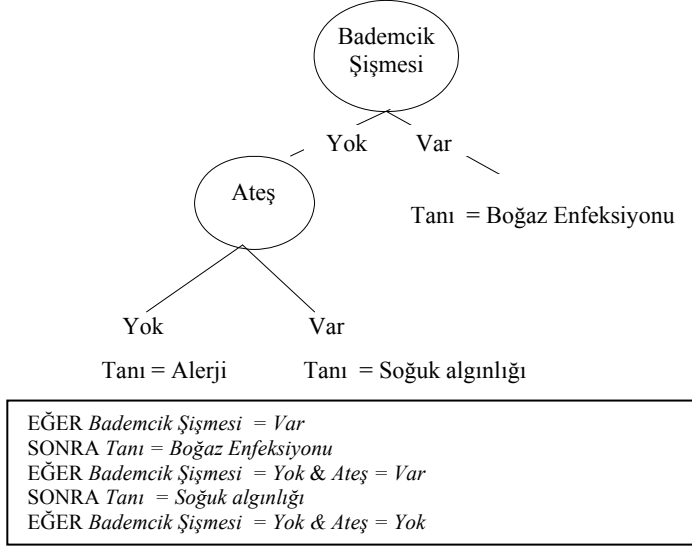
- Bölme: Bu işlem, verilerin daha küçük alt kümelere ayrılmasını içeren yinelemeli bir süreçtir. İlk yineleme tüm veriyi içeren kök düğümünü ele alır. Bundan sonraki yinelemeler, verinin alt kümelerini içeren türev düğümler üzerinde işlem yapmaktadır. Her bölme işleminde, değişkenler analiz edilir ve en iyi bölme seçilir.
- Budama: Bir ağaç oluşturulduktan sonra, istenmeyen alt ağaçlar veya düğümler içerebilir. Budama işlemi ile bunlar ayıklanır. Bir başka ifade ile budama işlemi bir karar ağacını daha genel yapmak için kullanılmaktadır.
- Durdurma kriteri: Ağaç oluşturma algoritmaları çeşitli durdurma kuralları içerirler. Bu kurallar genellikle, maksimum ağaç derinliği, bir düğümden bölme için ele alınan minimum eleman sayısı ve yeni bir düğümden bölme için ele alınan minimum eleman sayısı gibi çeşitli faktörlere dayanır.

Eğitim işlemi gerçekleştirildikten sonra, yeni bir veri örneği için elde edilen ağacın yardımı ile ağacın en tepesinden başlanarak ve bir yaprak düğümü ile karşılaşılıncaya kadar olan dalların oluşturduğu yol izlenerek tahmin yapılabilir. İzlenecek yol, yeni örnek içindeki bağımsız değişkenlerin değerleri üzerinde elde edilen karar ağacının bölme kuralları uygulanarak belirlenir (Bounsaythip ve Esa, 2001,20). Karar ağaçlarına örnek olarak, Tablo 1’de kategorik verileri içeren bir hasta veri tabanı, Şekil 1’de buna ait bir karar ağacı ve bu ağaçtan elde edilen kurallar verilmiştir.

Tablo 1: Bir Hasta Veritabanı

| HASTA SIRA NO.SU | BOĞAZ AĞRISI | ATEŞ | BADEMCİK ŞİŞMESİ | KAN TOPLAMASI | BAŞ AĞRISI | TANI |
|------------------|--------------|------|------------------|---------------|------------|-------------------|
| 1 | Var | Var | Var | Var | Var | Boğaz Enfeksiyonu |
| 2 | Yok | Yok | Yok | Var | Var | Alerji |

| | | | | | | |
|----|-----|-----|-----|-----|-----|-------------------|
| 3 | Var | Var | Yok | Var | Yok | Soğuk algınlığı |
| 4 | Var | Yok | Var | Yok | Yok | Boğaz Enfeksiyonu |
| 5 | Yok | Var | Yok | Var | Yok | Soğuk algınlığı |
| 6 | Yok | Yok | Yok | Var | Yok | Alerji |
| 7 | Yok | Yok | Var | Yok | Yok | Boğaz Enfeksiyonu |
| 8 | Var | Yok | Yok | Var | Var | Alerji |
| 9 | Yok | Var | Yok | Var | Var | Soğuk algınlığı |
| 10 | Var | Var | Yok | Var | Var | Soğuk algınlığı |

Şekil 1. Hasta Veri Tabanı İçin Bir Karar Ağacı ve Kurallar

4. KARAR AĞACI ALGORİTMALARI

Temelleri AID (Automatic Interaction Detector) yöntemi ile atılan karar ağacı modelleri çeşitli algoritmalar ile sürdürülmüştür. Morgan ve Sonquist adlı araştırmacılar tarafından 1970'li yılların başlarında önerilen ve kullanılan AID algoritması, karar ağacı tabanlı ilk algoritma ve yazılımdır. Bilgisayar biliminde veri grubunu bir karar ağacı ile tanımlama işlemi uygulanan bir yöntem olmasına rağmen, bu yöntem özbilgiyi elde etmede uzun yıllar tercih edilmemiştir.

1984 yılında Berkeley Üniversitesi'nden Leo Breiman ve Charles J. Stone ile Stanford Üniversitesi'nden Jerry Friedman ve R. Olshen tarafından basılan "Classification And Regression Trees" adlı kitapta yeni bir karar ağacı yordamı olan C&RT algoritmalarının kullanılmasından bahsedilmektedir. Bu çalışma, yöntemin istatistik biliminde yer edinmesini sağlamıştır. 1986 yılında J.R. Quinlan adlı araştırmacı karar ağaçlarına yeni bir algoritma eklemiştir. Bu karar ağacı algoritması literatüre ID3 algoritması olarak geçmiştir. 1993 yılında ise Quinlan adlı bir başka araştırmacı "Programs For Machine Learning" adlı kitabında C4.5 karar ağacı algoritmasını ortaya koymuştur (Lee ve Keng, 2001, 44).

Geliştirilen diğer algoritmalar arasında CHAID (G.V. Kass; 1980), Exhaustive CHAID (Biggs, de Ville ve Suen; 1991), MARS (Multivariate Adaptive Regression Splines; Friedman), QUEST (Quick, Unbiased, Efficient Statistical Tree; Loh ve Shih, 1997), C5.0 (Quinlan), SLIQ (Mehta, Agarwal ve Rissanen), SPRINT (Shafer,

Agrawal ve Mehta) yer almaktadır (Kirchner, 2004, 116). Tablo 2’de bazı karar ağacı algoritmalarının özellikleri verilmektedir (Bounsaythip ve Esa, 2001, 21):

Tablo 2. Bazı Karar Ağacı Algoritmaları Özellikleri

| KARAR AĞACI ALGORİTMASI | ÖZELLİKLER |
|---|--|
| C&RT | Gini'ye dayalı ikili bölme işlemi mevcuttur. Son veya uç olmayan her bir düğümde iki adet dal bulunmaktadır. Budama işlemi ağacın karmaşıklık ölçüsüne dayanır. Sınıflandırma ve regresyonu destekleyici bir yapıdadır. Sürekli hedef değişkenleri ile çalışır. Verinin hazırlanmasına gereksinim duyar. |
| C4.5 ve C5.0 (ID3 karar ağacı algoritmasının ileri versiyonları) | Her düğümde çıkan çoklu dallar ile ağaç oluşturur. Dalların sayısı tahmin edicinin kategori sayısına eşittir. Tek bir sınıflayıcıda birden çok karar ağacını birleştirir. Ayırma işlemi için bilgi kazancı kullanır. Budama işlemi her yapraktaki hata oranına dayanır. |
| CHAID (Chi-Squared Automatic Interaction Detector) | Ki-kare testleri kullanarak bölme işlemi gerçekleştirir. Dalların sayısı iki ile tahmin edicinin kategori sayısı arasında değişir. |
| SLIQ (Supervised Learning In Quest) | Hızlı ölçeklenebilir bir sınıflayıcıdır. Hızlı ağaç budama algoritması mevcuttur. |
| SPRINT (Scalable Parallelizable Induction of Decision Trees) | Büyük veri kümeleri için idealdir. Bölme işlemi tek bir niteliğin değerine dayanır. Tüm bellek sınırlamaları üzerinde nitelik listesi veri yapısı kullanarak işlem yapar. |

Kaynak: Bounsaythip, Catherine ve Esa Rinta-Runsala, "Overview of Data Mining For Customer Behavior Modeling", VTT Information Technology Research Report, Version:1, 2001, s. 21.

5. C&RT İLE BİR SATIŞ ANALİZİ UYGULAMASI

5.1. Uygulamanın Amacı

Burada; müzik CD'leri, küçük ev eşyaları, alet, oyuncak, hediyelik eşya, oyun, magazin, video, cep telefonu ve cep telefonu aksesuarları satışı yapan bir perakendeci işletmenin satış analizi ele alınmaktadır. İşletme, daha çok doğrudan pazarlama yapmakta ve satışlarının küçük bir kısmını yerinde satış, geriye kalanını ise posta, telefon, faks ve internet (e-mail) yöntemleri ile sipariş alarak gerçekleştirmektedir. Analizin amacı; işletmenin gerçekleşen toplam satış hasılatının ürünlere göre dağılımının ayrıntılarını ve ürünlerin toplam satış değerlerinin toplam satış hasılatı üzerindeki göreceli önemini, müşteriler ile ilişkilendirerek elde etmektir. Bir başka ifade ile, birden fazla sipariş yöntemi kullanabilen müşterilerin harcama örüntülerini belirleyerek, satış değeri üzerindeki daha karlı ya da öncelikli

hedef müşteri gruplarını ve aynı zamanda yeni gelecek müşterilerin sınıfını belirleyebilen bir tahmin modeli oluşturmaktır.

5.2. Uygulamanın Kapsamı ve Yöntemi

Bu çalışmada, yapılan analiz için C&RT karar ağacı tekniği kullanılmakta ve bunun için de Clementine v8.1 veri madenciliği yazılımından yararlanılmaktadır. Clementine programı veri akışına dayanmaktadır. Bu akış içinde düğümler ile gösterilen her bir işlem, birbirine oklar ile bağlanmaktadır. Buradaki analiz için oluşturulan veri akışı Ek 1’de verilmektedir.

Analiz, belirli bir dönemde satın alınan ürünleri ve tutarlarını, yerinden ve doğrudan satış yöntemleri ile yapılan satın alma sıklıklarını içeren işletmenin 8000 adet müşteri kaydından oluşan bir müşteri veri tabanı üzerinde gerçekleştirilmektedir. Veri tabanında işletmenin satış verileri müşteriye göre kişiselleştirilmiştir. Müşteri kayıtlarında yer alan değişkenler Tablo 3’te verilmektedir:

Tablo 3. Müşteri, Satış ve Satış Yöntemleri Nitelikleri Kayıtları

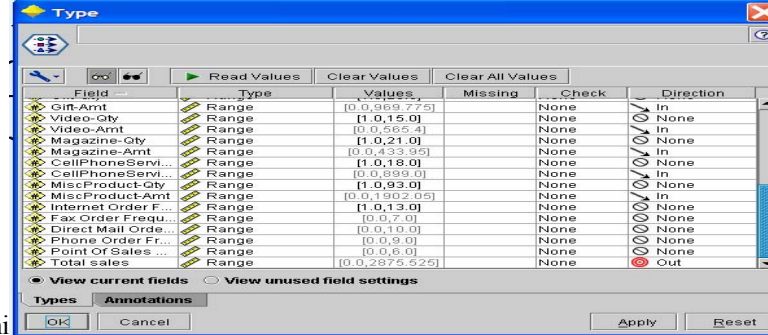
| DEĞİŞKENLER | DEĞİŞKENLERİN AÇIKLAMASI |
|--------------------------|--|
| Customer_id | Müşteri Numarası |
| Customer status | Müşteri Durumu |
| City name | Bulunduğu Şehir |
| State code | Eyalet Kodu |
| Zip code | Zip Kodu |
| Book_Qty | Alınan Kitap Adedi |
| Book_Amt | Kitaplara Ödenen Tutar |
| Music CDs-Qty | Alınan CD Adedi |
| Music CDs-Amt | CD'lere Ödenen Tutar |
| Housewares-Qty | Alınan Ev Aletlerinin Adedi |
| Housewares-Amt | Ev Aletlerine Ödenen Tutar |
| Tools-Qty | Alınan Alet Adedi |
| Tools-Amt | Aletlere Ödenen Tutar |
| Games-Qty | Alınan Oyun Adedi |
| Games-Amt | Oyunlara Ödenen Tutar |
| Gift-Qty | Alınan Hediye Adedi |
| Gift-Amt | Hediyelere Ödenen Tutar |
| Video-Qty | Alınan Video Adedi |
| Video-Amt | Videolara Ödenen Tutar |
| Magazine-Qty | Alınan Magazin Adedi |
| Magazine-Amt | Magazinelere Ödenen Tutar |
| CellPhoneService-Qty | Alınan Cep Telefonu Servisi |
| CellPhoneService-Amt | Cep Telefonu Servisi İçin Ödenen Tutar |
| Internet Order Frequency | İnternet İle Alışveriş Sıklığı |
| Fax Order Frequency | Fax Yolu İle Alışveriş Sıklığı |

| | |
|--------------------------------|------------------------------------|
| Direct Mail Order Frequency | Posta Yolu İle Alışveriş Sıklığı |
| Phone Order Frequency | Telefon Yolu İle Alışveriş Sıklığı |
| Point of Sales Order Frequency | Yerinde Satış Sıklığı |
| Total Sales | Toplam Satışlar (Harcama Tutarı) |

5.3. Analiz

Ele alınan analiz, toplam satışların üzerinde hangi satış faktörlerinin başka bir ifade ile ürün satışlarının ve müşteri satın alma örüntülerinin etkilerini belirlemedir. Analize Ek 1’de görüldüğü gibi verilerin programa yerleştirilmesi ile başlanır. Değişkenlerin tipini belirlemek için ise “Type” düğümü kullanılır. Bu düğüm yardımı ile çıktı (target) olarak toplam satışlar, girdi (input) olarak da ürünlere ödenen tutarlar seçilir. Bu seçim işlemi Şekil 2’te gösterilmektedir.

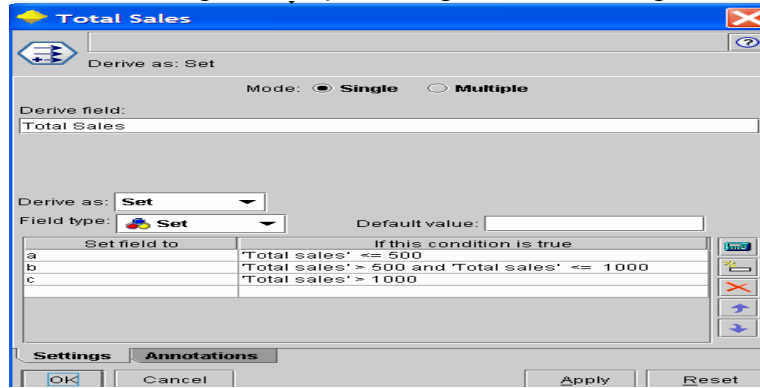
Şekil 2: Toplam Satışların Analizi İçin “Target” ve Input”



Seçimi

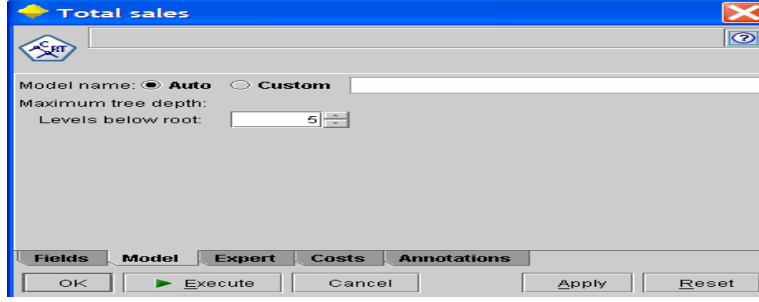
Toplam satışlar, sürekli bir değişken olup işletme tarafından önemli görülen seviyeler olan a; 500 pb ve altı, b; 500 pb üstü ile 1000 pb arası, c; 1000 pb üstü olmak üzere üç sınıfa ayrılmıştır. Bu işlem için “Type” düğümüne “Derive” düğümü bağlanmaktadır. Bu değerlerin “Derive” düğümünde programa yerleştirilmesi Şekil 3’de gösterilmektedir.

Şekil 3: Toplam Satışlar İle İlgili “Derive” Düğümü



Oluşturulan “Derive” düğümüne 5 seviyeli C&RT karar ağacı düğümü bağlanmıştır. Bununla ilgili program penceresi Şekil 4’de verilmektedir. Elde edilen karar ağacının sonuçları ise Ek 2’de gösterilmektedir.

Şekil 4: Beş Seviyeli Karar Ağacı Düğümü



Oluşturulan bu karar ağacına göre; 500 pb ve altında alışveriş yapan 6326 kişiden 1751’i, 60.34 pb ve üstünde kitap satın almış, bu kişilerin de 124’ü 95 pb ve üstünde harcama yaparak alet satın almıştır. Benzer şekilde, 500 pb ve altında alışveriş yapan 6326 kişiden 4575’i ise 60.34 pb altında kitap satın almış, bu kişilerin de 190’ı 103.75 pb ve üstünde harcama yaparak alet satın almış ve alışverişlerini tamamlamıştır. Ek 2’de yer alan karar ağacı sonuçlarından yararlanarak benzer bir çok örüntü elde edilebilir.

5.4. Sonuçların Değerlendirilmesi

Gerçek değerler ile tahmin edilen değerler arasındaki yüksek korelasyon (1’e yakınlık) tahminin başarısı için önemlidir. Kullanılan C&RT düğümlerinin tahmin güvenilirliği değerleri, C&RT düğümüne “Analysis” düğümü eklenerek bulunmuştur. Toplam satışlar için istatistiki sonuçlar Şekil 5’de verilmektedir.

Şekil 5: Toplam Satışlar İçin “Analiz” Düğümü Sonuçları

| Results for output field Total sales | |
|---|-----------|
| Comparing \$X: Total sales with Total sales | |
| Minimum Error | -609.272 |
| Maximum Error | 1.261.628 |
| Mean Error | -0.000 |
| Mean Absolute Error | 65.320 |
| Standard Deviation | 100.684 |
| Linear Correlation | 0.950 |
| Occurrences | 8000 |

Burada toplam satışlar ile tahmin edilen toplam satışlar arasında yüksek korelasyon (0,95) mevcuttur. Bu da, yüksek tahmin başarısını göstermektedir.

Aykırı değerlerin belirlenip çıkarılması ve karar ağacı seviyesinin artırılması ile daha da yüksek tahmin başarısı elde edilebilir.

Müşterilerin harcama tutarına göre oluşturulan ve Ek 2’de verilen yüksek güvenilirlikli karar ağacı sonuçları incelendiğinde, işletme müşterilerinin %20.92’sinin toplam satışların yaklaşık %63.36’sını, bu grubun %2.2’sinin ise toplam satışların yaklaşık %14’ünü oluşturduğu görülmektedir. Bu durumda, c grubunda yer alan müşteriler en çok gelir getiren müşteri grubudur. Toplam müşterilerin küçük bir oranını oluşturmalarına rağmen, işletmeye sağladıkları gelir bakımından ilk sırada yer almaktadırlar. Müşterilerin %54.81’nin tahmin edilen harcaması 81.09 pb’dir. Harcamaları işletmenin toplam satışlarının sadece % 17.34’ünü oluşturan bu grup, 60.34 pb altında kitap, 103.75 pb nin altında ise alet satın almıştır. Yine, toplam müşterilerin yaklaşık %23,3’ü toplam satışların sadece %3,6’sını sağlamaktadır. Bu grup; oyundan 130.225 pb altında, kitaptan 25.96 pb altında ve aletten de 103.75 pb altında satın alma davranışı örüntüsüne sahiptir.

Yukarıda elde edilen satın alma davranışı örüntüleri, işletmenin pazar bölümlendirme, hedef pazar seçimi ve pazarlama karması stratejileri için karar destek aracı olarak kullanılabilir. Genel olarak bakıldığında, işletme müşterilerinin %20.92’si, toplam satışların %63.36’sını oluşturması nedeni ile önemli bir pazar bölümüdür. Özellikle, toplam satışların % 14’ünü oluşturan müşteri grubu, işletmenin öncelikli hedef müşteri grubu olup bir niş pazardır. Dolayısıyla; ürün, fiyat, dağıtım ve tutundurma stratejileri açısından, diğer müşteri gruplarının oluşturduğu pazar bölümlerinden daha farklı pazarlama stratejilerinin uygulanması gerekmektedir. Bunun dışında, toplam satışların %36.64’sını oluşturan % 79.08’lik diğer bir müşteri grubu mevcuttur. Bu pazar bölümü için de, satış düşüklüğünün nedenleri araştırıldıktan sonra, ürün çeşitlendirme, pazar geliştirme ve bazı satış artırıcı tutundurma stratejileri uygulanabilir. Bir başka ifade ile söz konusu analiz; işletmeye, yeni hedef pazar belirleme, mevcut pazarlara nüfuz etme ve mevcut pazarlardan çekilme gibi stratejileri için bir karar destek aracı olacaktır.

Elde edilen karar ağaçlarının yardımı ile daha birçok yararlı çıkarım ve örüntü elde etmek mümkündür. Ayrıca, buradan elde edilen sonuçlar başka ayrıntılı analizlerin yapılmasına da neden olabilecektir.

6. SONUÇ

Günümüzün tüketici odaklı pazarlarında işletmeler süreklilik arz eden yoğun bir rekabetin içindedirler. İşletmelerin bu rekabet şartlarında başarılı olabilmeleri için etkin ve düşük maliyetli pazarlama stratejileri uygulamaları gerekmektedir. Etkin pazarlama stratejilerinin oluşturulabilmesi için doğru bilgilere, doğru bilgilerin elde edilebilmesi için ise verileri çok boyutlu analiz edebilen veri madenciliği gibi araçlara gereksinim duyulmaktadır.

Bu çalışmada, bir satış hacmi analizi için veri madenciliğinin bir sınıflama işlevi olan karar ağacı yöntemi uygulanmıştır. Mutlak ölçümler yapan makro satış hacmi analizleri günümüzde yeterli olmayıp satış bölgesi, tüketici tipi veya ürün kategorisi gibi alt alanları içerdiğinde daha anlamlı olmakta ve işletmeler tarafından istenilen

göreceli değerleri ölçer hale gelmektedir. Burada, ele alınan işletmenin belirli bir dönemde yaptığı satışlar müşteri satın alma davranış özelliklerine dayalı olarak analiz edilmiştir. Yapılan C&RT karar ağacı tekniği uygulaması sonucunda harcama tutarına göre müşteriler sınıflara ayrılmıştır.

Bu şekilde, satış başarımında oluşan hedef boşluklarının ve buna farklı faktörlerin göreceli katkısının ne olduğu ortaya konabilmektedir. Müşterilerin tanımlanması ve hedef müşteri grubunun belirlenmesi, hatta belirlenen hedef pazarların adlandırılabilmesi, özel teklif sunmaya geçecek müşteri veya adayların seçilmesine, müşterilerin işletmeye bağlılığının güçlendirilmesine ve tekrar satın almalarının sağlanmasına veya müşteri portföyünün değiştirilmesine yönelik etkin pazarlama stratejilerinin oluşumuna önemli katkıda bulunmaktadır.

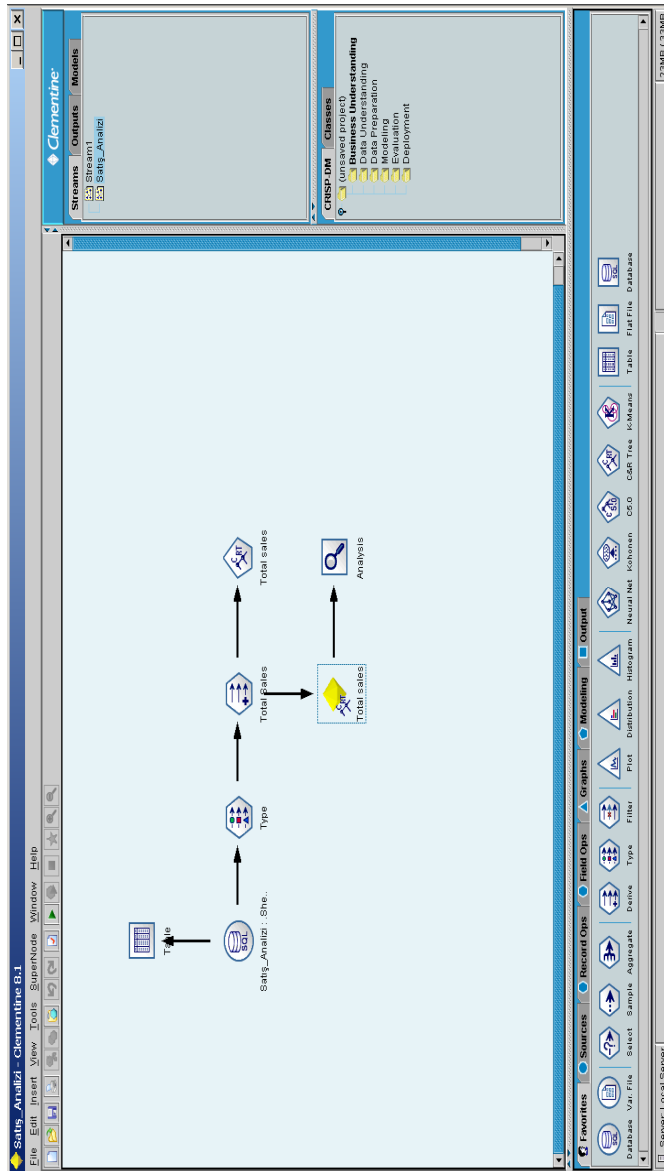
Ayrıca, elde edilen sağlıklı bilgilere dayalı pazarlama stratejileri envanter, dağıtım, müşteriye hızlı cevap verebilme gibi işletmenin bir çok politikasını olumlu yönde etkilemektedir. Böylece maliyetlerin düşürülmesi sureti ile kar artışı meydana gelebilmektedir. Çok güçlü bir araç olan veri madenciliği işletmelerin ihtiyacı olan bu tip çok çeşitli bilgiyi çok hızlı ve kolay bir şekilde sunacağından, önümüzdeki yıllarda rekabetçi ve proaktif işletmelerin vazgeçilmez bir karar destek aracı haline gelmesi kaçınılmazdır.

Kaynaklar

- Bloemer, M. M. J., Brijs, T., Vanhoof, K. ve Swinnen, G. (2003); "Comparing Complete And Partial Classification For Identifying Customers At Risk", *International Journal Of Research In Marketing*, Vol:20, Issue:2, s. 117-131.
- Bounsaythip, C. ve Esa, R. R. (2001). "Overview of Data Mining For Customer Behavior Modeling", *VTT Information Technology Research Report*, Version:1, s. 1-53.
- Carvalho, D.R. ve Freitas, A.A. (2004), "A Hybrid Decision Tree/Genetic Algorithm Method For Data Mining", *Information Sciences*, 1-3(163), s. 13-35.
- Chen, J. (2001). "A Predictive System For Blast Furnaces By Integrating A Neural Network With Qualitative Analysis", *Engineering Applications of Artificial Intelligence* 1(14), s. 77-85.
- Clementine User's Guide. (2002).
- Gargano, M.L. ve Raggad, B.G. (1999). "Data mining—A Powerful Information Creating Tool", *OCLC Systems & Services* 2(15), s. 81–90.
- Ge, A. X. (1999). A Neural Network Approach to the Modeling of Blast Furnace, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, (Unpublished Master Thesis), <http://scanner-group.mit.edu/PDFS/RaviT.pdf>.
- Groth, R. (2000). *Data Mining: Building Competitive Advantage*, Prentice-Hall, USA.

- Kirchner, K., Tölle, K. –H. ve Krieter J.(2004). The Analysis Of Simulated Sow Herd Datasets Using Decision Tree Technique, *Computers and Electronics In Agriculture*, 2(42), s. 111-127.
- Lee, S. J. ve Keng, S. (2001). “A Review Of Data Mining Techniques”, *Industrial Management&Data Systems I*(101), s. 41-46.
- Liao, S. (2003). “Knowledge Management Technologies And Applications— Literature Review From 1995 To 2002”, *Expert Systems With Applications* 2(25), s. 155–164.
- Ma, Y. (1998). Data Warehousing, OLAP, And Data Mining: An Integrated Strategy For Use At FAA, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, <http://www.citeseer.ist.psu.edu/ma98data.html>.
- Metaxiotis, K., Ergazakis, K., Samouilidis, E. ve Psarras J., (2003). “Decision Support Through Knowledge Management:The Role Of The Artificial Intelligence”, *Information Management & Computer Security* 5(11), s. 216-221.
- Moshkovich, H.M., Mechitov, A.I. ve Olson, D.L. (2002). “Rule Induction In Data Mining: Effect Of Ordinal Scales”, *Expert Systems With Applications* 4(22), s. 303-311.
- Rygielski, C., Wang, J.C. ve Yen, D.C. (2002). “Data mining Techniques For Customer Relationship Management”, *Technology in Society* 4(24), s. 483-502.
- Ryu, T. W. ve Eick, C. F. (2004). A Database Clustering Methodology and Tool, *Information Sciences*, Basım Aşamasında.
- Shaw, M.J., Subramaniam, C, Tan, G.W. ve Welge M.E. (2001). “Knowledge Management And Data Mining For Marketing”, *Decision Support Systems I*(31), s. 127–137.
- Kotler, P. (2002). *Marketing Management*. Prentice-Hall, Milenium Edition, USA.
- Yun, H., Ha, D., Hwang, B. ve Ryu, K.H. (2003); “Mining Association Rules On Significant Rare Data Using Relative Support”, *Journal of Systems and Software*, Vol:67, Issue:3, s.181-191. www.spss.com/classification_trees.
- Zorman, M., Vili, P., Kokol, P., Peterson, M., Sprogar, M. ve Ojstersek, M. (2001). “Finding The Right Decision Tree’s Induction Strategy For A Hard Real World Problem”, *International Journal Of Medical Informatics I-2*(63), s. 109-121.

Ek 1:



Ek 2:

