Istanbul University Journal of the School of Business
İstanbul Üniversitesi İşletme Fakültesi Dergisi

Vol/Cilt: 46, No/Sayı:1, May/Mayıs 2017, 62-75
ISSN: 1303-1732 – http://dergipark.ulakbim.gov.tr/iuisletme

Web Günlük Dosyalarının Analizi için Web Kullanım Madenciliğinin Uygulanması

Applying Web Usage Mining for the Analysis of Web Log Files

Serra Çelik[a,✦]

YAYIN BİLGİSİ
Article Info

Başvuru/Received
23/06/2016

Kabul/Accepted
14/03/2017

Anahtar Sözcükler:

Web Kullanım Madenciliği
Destek Vektör Makineleri
İnternet
Sınıflandırma

Keywords:

Web Usage Mining
Support Vector Machines
Internet
Classification

ÖZ

Günümüzde veri artışı inanılmaz boyutlara ulaşmıştır. Gelişen teknolojiyle birçok farklı sektörde daha kolay veri elde edilebilmektedir. Bu noktada veri madenciliği bu veri yığınlarından anlamlı bilgiye dönüşüm sürecini hızlandırmıştır. Veri madenciliği, ilk başta veri tabanlarından bilgi çıkarımı olarak ortaya çıksa da günümüzde geliştirilen yeni yöntemler ve teknolojilerin desteği ile tahmin gücünden daha fazla yararlanılmaktadır. Çalışmada veri madenciliği sınıflandırma yöntemlerinden destek vektör makineleri, web kullanım madenciliği verisi olan web günlük dosyaları üzerine uygulanmıştır. Kullanılan veri seti bir e-ticaret sitesinin 812 güne ait web günlük dosyalarıdır. Web günlük dosyaları yapılandırılmamış veri içermektedir ve bu tip verinin analizi yapılandırılmış veriye göre daha zordur. Bu nedenle analiz öncesinde verinin temizlenmesi gerekmiş ve bu süreç çalışmada uzun bir süre almıştır. Çalışmada satın alma davranışının eğilimini belirlemek hedeflenmiştir. Destek vektör makineleriyle sınıflandırma yapılmış sonuçlar lojistik regresyonla elde edilen sonuçlarla karşılaştırılmıştır. Destek vektör makineleri ile bir e-ticaret sitesi uygulamasında daha doğru sınıflandırma yapılabildiği görülmüştür.

ABSTRACT

Today, size of data has reached amazing amounts. Recent advances in technology collecting data in many different sectors is getting easier. At this point, data mining has accelerated the process of transforming data to information. In the beginning, data mining has been known as information extraction from databases, but recently it is more useful for prediction by the help of new methods and technologies developed. In this study web usage mining will be performed with classification methods of data mining using web log files. The data used is an e-commerce web site's log files of 812 days. Web log files contain unstructured data and it is very difficult to analyze it in conventional ways. Before analyzing the data, it has to be cleaned and this process takes long time. The aim of this study is finding the way of purchase behavior. First, analysis is made by support vector machines, then results are compared with the results obtained by logistic regression. For implementation to an e-commerce web site, it can be stated that support vector machines can classify more accurately.

✦ İlgili yazar. Tel: +90 (212) 473 7070 – 11538, e-posta: serra.celik@istanbul.edu.tr (S. Çelik)

[a] *Enformatik Bölümü, İstanbul Üniversitesi, İstanbul, Türkiye*

Bu çalışma, Dr. Serra Çelik tarafından, Prof.Dr. M. Erdal Balaban danışmanlığında hazırlanan "Web Kullanım Madenciliği ile Satın Alma Davranışının Modellenmesi ve Bir E-Ticaret Uygulaması" başlıklı doktora tezinden derlenmiştir.

1. Giriş

İnternetin artık yaşamın vazgeçilmez bir unsuru olması, e-ticaretin hızlı gelişimi ve başarısı, pazarlama yöneticilerinin karar almalarını kolaylaştırıcı modelleri kullanmalarını gerektirmektedir. Son dönemde internet ve e-ticaret üzerine yapılan modelleme çalışmaları; geleneksel perakendecilik, fiyatlama, promosyonlar, müşteri hizmetleri, satın alma davranışı ve müşteri sadakati üzerine yoğunlaşmıştır. Elektronik ticaret yapan şirketler web sitelerine potansiyel müşterileri (ziyaretçileri) çekmek, satın alma davranışını anlamak ve müşterilerinin sürekliliğini sağlamayı amaçlamaktadırlar (Bucklin, 2008).

Şirket kârlılığı için satışlar ve müşteri potansiyelinin tanımlanması kritik öneme sahiptir. Bu alan doğrudan, etkileşimli, hedef kitle ve veri tabanı pazarlama olarak bilinmekte ve pazarlama alanında araştırmacı ve uygulamacılar için büyük önem arz etmektedir (McCarty ve Hastak, 2007). Buradaki amaç, satın alım gibi müşterinin gelecek davranışının tahmini için müşterinin işlem ve davranışsal verisinin analiz edilmesidir (Hughes, 2005).

Şirket veri tabanlarının etkin kullanımı ve veri analiz yöntemlerindeki gelişmelerle müşteri veri analizi de önemli bir hale gelmiştir. Pazarlama çalışmalarında analiz yöntemi olarak karar ağaçları, yapay sinir ağları, lojistik regresyon analizi gibi makine öğrenmesi yöntemlerini içeren müşteri tepki modelleri kullanılmaktadır (Bose ve Chen, 2009).

Müşteri tepki modellerinde "sınıf dengesizliği" gerçek uygulamalarda da ortaya çıkan önemli bir sorundur. Ancak literatürde bu problemin yeterince ilgi görmediği görülmektedir. Örneğin, satın alım olması durumu için "1", satın alım olmaması durumu için "0" değerinin atandığı iki kategorili sınıflandırma modelinin veri setinde "1"lerin yani satın alımların "0"lara oranla oldukça az olduğu gözlenmektedir. Sigorta dolandırıcılık tespiti, petrol sızıntısı tahmini, erken dönem doğum tahmini, sistem başarısızlık tahmini, afet tahmini gibi örnekler bu tür verilere örnek verilebilmektedir (Ling ve Li, 1998), (Weiss, 2004). Bu veri setlerinde veri analizi yöntemleri çoğunluk durumlarına göre taraflı davranma eğilimi gösterebilmekte ve azınlık durumları da çoğunluğa göre sınıflandırılabilmektedir. Bu problem müşteri tepki tahmini (McCarty ve Hastak, 2007) ve müşteri kayıp tahminlerinde (Burez ve Van den Poel, 2009) sıklıkla görülmektedir. McCarty ve Hastak (2007) %3'ten, Ling ve Li (1998) ise %1,5'tan daha az tepki oranına sahip veri setleriyle çalışmışlardır. Dengesizlikle uygun bir şekilde başa çıkılamazsa model, gerçek tepki oranının zayıf bir tahminini sunmaktadır. Daha açık bir ifadeyle, modelin tahmin doğruluğu artmakta ancak ilgi sınıfı (satın alma, sahtekârlık, ikna, vb.) hakkında yanlış bilgi vermektedir. Bu doğruluk oranının da, yanıltıcı olduğu sonucu ortaya çıkmaktadır.

E-ticaret sitelerinde ziyaretçiler her zaman satın alım kararı ile web sitesini ziyaret etmemekte; ürün kıyaslama, fiyat öğrenme gibi amaçlarla da sitelere erişmektedir. Bu durum satın alım sayısının ziyaret sayısından oldukça küçük olmasına, başka bir ifadeyle müşteri tepki oranının düşük kalmasına yol açmaktadır. Bu çalışmada kullanılan e-ticaret web sitesi verisi için bu oran yüzde birden daha küçüktür. Literatürde, müşteri tepki modelleri çerçevesinde çok sayıda uygulama olsa da e-ticaret alanında web günlük verileriyle yapılmış çalışmaya rastlanmamıştır. Makine öğrenmesi yöntemleri ise müşteri tepki modellerinde oldukça seyrek kullanılmakta, ancak bu çalışma kapsamında bir çalışmaya rastlanılmamıştır. Bu çalışmada, makine öğrenmesi yöntemlerinden Destek Vektör Makineleri (DVM)'nin web kullanım madenciliği kapsamında müşteri tepki modeli olarak kullanılabilirliği test edilmiştir. Tepki modellerinde sıkça kullanılan çok değişkenli güçlü bir istatistik yöntem olan lojistik regresyon ile de sonuçlar karşılaştırılmıştır.

DVM'de dengesiz veri durumunda karar sınırını azınlık sınıfına kaydırır. Çoğu veri noktası, çoğunluk sınıfında sınıflanır. Bu da modelin zayıf olmasına neden olabilmektedir.

Sınıf dengesizliğiyle başa çıkabilmek için kullanılan başlıca yöntemler sınıflandırıcı değişimi (classifier change) ve yeniden örnekleme (resampling) yöntemleridir (Weiss, 2004). Sınıflandırıcı değişim modelinde taraflı tahminden kaçınmak için amaç fonksiyonu üzerinde farklı ağırlığa sahip tahminleyiciler oluşturulur (He ve Garcia, 2009). Yeniden örnekleme yönteminde ise, veriyi dengelemek için veri arttırılır (oversampling) ya da azaltılır (undersampling) (He ve Garcia, 2009), (Japkowicz, 2001).

Veri setini azaltan örnekleme (örnek küçültme) yönteminde veriyi dengelemek için çoğunluk sınıf verisi, tesadüfi olarak azınlık sınıfına eşit ya da katları olarak alınmaktadır. Yöntem, veri setinde sınıf dengesini kolaylıkla sağlayabilmektedir (Kim, Chae ve Olson, 2013).

Veri setini arttıran örnekleme (örnek büyütme) yönteminde ise azınlık sınıfından kopyalar üretilerek veri dengelenmek istenmekte, ancak bu da aşırı uyuma (overfitting) neden olabilmektedir (Chawla v.d., 2002), (Drummond ve Holte, 2003).

Japkowicz (2000), örnek büyütme ve örnek küçültme yöntemlerini değişik veri setleri için değerlendirmiş ve iki yöntemin de etkili olduğu sonucuna varmıştır. Örnek küçültme yöntemi, örnek büyütme yöntemine göre uygulanması daha kolay ve büyük veri setleri için daha

uygun kabul edilmektedir (Drummond ve Holte, 2003), (Khoshgoftaar, Van Hulse ve Napolitano, 2011). Chawla v.d. (2002), Synthetic Minority Oversampling Technique (SMOTE) adıyla bir yöntem geliştirmişlerdir. Bu yöntem, hayalet dönüşümü ile yeni özellikler oluşturmaktadır. Her pozitif örnek için onun en yakın pozitif komşusu tanımlanmakta ve yeni pozitif örnekler oluşturularak, komşuları arasında tesadüfi olarak yer almaktadır. Burada problem, öğrenmiş sınırın pozitif özelliklere oldukça yakın olmasıdır (Provost ve Fawcett, 2001).

Ngai, Xiu ve Chau (2009) çalışmalarında yapay sinir ağları, karar ağaçları ve regresyon analizinin müşteri ilişkileri yönetiminde geniş olarak kullanıldığını göstermiştir. Bu yöntemler müşteri segmentasyonu oluşturma ya da tepki modelinde de kullanılmaktadır (Bose ve Chen, 2009). Öte yandan Müşteri İlişkileri Yönetimi (MİY) ve müşteri tepki modeli için DVM kullanımı çok seyrektir (Viaene v.d., 2001).

Çalışmada Web günlük verileri üzerine Destek Vektör Makineleri (DVM) ve Lojistik Regresyon uygulanmıştır. Öncelikle Web Kullanım Madenciliği hakkında bilgi verilmiş, ardından Destek Vektör Makineleri açıklanmıştır. Son olarak da sınıflandırma yöntemi DVM ve karşılaştırma yöntemi olarak kullanılan Lojistik regresyon ile farklı modeller denenerek satın alım sınıflandırma sonuçları değerlendirilmiştir.

2. Web Kullanım Madenciliği

WWW'nun hızla gelişmesiyle internet, faydalı bilgilerin bulunabileceği bir kaynak olmuştur. Bu kaynak aynı zamanda bir veri kaynağıdır. Kullanıcılar, internette sayfalarda gezinirken, ürün/hizmet satın alırken, sosyal medyada paylaşımda bulunurken izler bırakmaktadırlar. Bu izler web sunucularında depolanmakta ve karşımıza yapılandırılmamış veri olarak çıkmaktadır. Web üzerinde üretilen ve depolanan bu veri web madenciliği ile analiz edilebilmektedir. Web madenciliği, web verisine veri madenciliği yöntemlerinin uygulanmasıdır. Daha açık bir ifadeyle web verisinden kullanıcı örüntüleri keşfetmede veri madenciliği yöntemlerinin uygulanması olarak tanımlanabilmektedir (Srivastava v.d., 2000). Web madenciliği terimi ilk kez Etzioni (1996) tarafından kullanılmıştır. Web kullanım madenciliği ise ilk kez Cooley v.d. (1997) tarafından, web sunucularından kullanıcı erişim örüntülerinin otomatik keşfi olarak tanımlanmıştır.

Web madenciliğinin en çok kullanılan sınıflama yöntemleri, Kosala ve Blockeel (2000) tarafından önerilmiş olan web'in hangi alanına madencilik yapılacağına ilişkin, üç ana kategoriye ayrılan sınıflamadır. Bu üç kategori; web içerik madenciliği, web yapı madenciliği ve web kullanım madenciliğidir.

Web içerik madenciliği web içerikleri ve belgelerinden kullanışlı bilginin keşfi olarak tanımlanmaktadır (Kosala ve Blockeel, 2000). Örneğin, web içerik madenciliğiyle web sayfaları konularına göre otomatik olarak kümelenip sınıflandırılabilmektedir. Ürün tanımları, forum paylaşımları gibi verilerin çıkarımıyla müşteri görüşleri irdelenip tüketici hassasiyeti keşfedilebilmektedir (Jiawei, Kamber, ve Pei, 2011), (Liu, 2007). Web yapı madenciliği bağlantılardan (hyperlinks), web yapısını temsil eden kullanışlı bilginin keşfi olarak tanımlanmaktadır (Örneğin, ortak ilgileri olan kullanıcı topluluklarının keşfi) (Xu v.d., 2011). Web kullanım madenciliğinde hedef, bir web sitesiyle ilişkili kullanıcı profillerini ve davranışsal örüntüleri analiz etmek ve modellemektir (Liu, 2007). İnternetin gelişmesi mevcut bilginin yayılmasına öncülük etmiş ve bu bilginin kişiselleştirilmesi ihtiyaç haline gelmiştir (Batista ve Silva, 2001).

Web kullanım madenciliğinin veri kaynağını oluşturan web günlüklerinin analizleri, web site yöneticilerinin yeterli bant genişliği ve sunucu kapasitesi sağlamada bir yol olarak sunulmasıyla başlamıştır. Bu analiz alanı, geçen sürede büyük gelişmeler yakalamıştır. E-şirketler, web günlük dosyalarını ziyaretçi profilleri ve satın alım faaliyetleri hakkında bilgi edinmek amaçlı kullanmaya başlamışlardır (Agosti ve Di Nunzio, 2007). Geçmiş web günlüklerinden ortaya çıkartılmış örüntüler ve mevcut gezinti örüntülerinin analiziyle kullanıcı davranışını tahmin etmek mümkündür. Kullanıcıların davranışını keşfetme, kişiselleştirme, sistem geliştirme ve kullanıcı ilgilerine göre sistem tasarımı, bir web sitesi modifikasyon sürecidir (Chitraa ve Davamani, 2010).

Web kullanım madenciliği için Srivastava v.d. (2000)'nin önerdiği süreç dört aşamadan oluşmaktadır. Girdi (input) aşaması, ön işleme aşaması (preprocessing stage), örüntü keşfi aşaması (pattern discovery stage), örüntü analizi aşaması (pattern analysis stage).

1. Girdi: Girdi aşamasında erişim günlükleri (access logs), referans günlükleri (referrer logs), vekil günlükleri (agent logs) olmak üzere üç tip ham web günlük dosyası kullanılmaktadır.

2. Ön İşleme: Ham web günlükleri veri madenciliğine olanak sağlayıcı bir biçimde değildir. Bundan dolayı ön işleme aşaması en önemli aşamalardan biridir. En yaygın kullanılan veri ön işleme görevleri (1) veri temizleme ve filtreleme, (2) robottan arındırma (de-spidering), (3) kullanıcı tanımlama, (4) oturum tanımlama, (5) yol tamamlamadır.

3. Örüntü Keşfi: Ön işleme tamamlandığında web verisi istatistiksel uygulamalar ve veri madenciliği yöntemleri için hazır hale gelmektedir. Bu yöntemler (1) standart istatistiksel analiz, (2) kümeleme algoritmaları, (3)

birliktelik kuralları, (4) sınıflandırma algoritmaları, (5) ardışık örüntülerdir.

4. Örüntü Analizi: Örüntü keşif aşamasında meydana çıkarılan örüntülerin hepsi kullanışlı olmayabilir. Amaca uygun olarak seçilmelidir.

Veri madenciliği uygulamalarında önemli bir adım, veri madenciliği ve istatistiksel algoritmaların uygulanabileceği uygun hedef veri setinin oluşturulmasıdır. Bu, tıklama akışı verisinin kendine has özelliğinden ve farklı kaynaklardan toplanmış veriyle ilişkili olduğundan web kullanım madenciliğinde oldukça önemlidir. Bu sürece veri hazırlama denilmektedir. Veri hazırlama süreci web kullanım madenciliğinde en fazla zaman harcayan ve yoğun hesaplama gerektiren aşamadır. Süreç, veriden kullanışlı örüntülerin başarılı şekilde çıkarımı için kritik öneme sahiptir (Liu, 2007). Bir e-ticaret sitesinin veri analizinde ön işleme yöntemlerinin doğru uygulanması, kullanıcı ve site metriklerini ortaya çıkarıcı öneme sahiptir (Kohavi v.d., 2004).

Ön işleme; veri kaynaklarından toplanan, örüntü keşfi için gerekli veriyi düzenleme sürecidir. Uç değerler, hatalar ve web taramadan kaynaklı meydana gelebilen tamamlanmamış veri tespit edilmektedir. Sunucu günlüklerinde kaydedilen veri, kullanıcıların web sitesine erişimlerini yansıtmaktadır. Vekil (agent) ve IP adresleri, kullanıcıları ve oturumlarını tanımlamaktadır. Bununla birlikte bazı sayfa görüntülemeleri kullanıcının tarayıcısında ya da vekil sunucusunda önbelleklenmiş olabileceğinden, sunucu günlükleriyle toplanan verinin tümüyle güvenilir olamayacağı gözden kaçmamalıdır. Bir web sunucu günlüğünde bir vekil sunucusundan tüm istekler aynı tanımlayıcıya sahiptir. Web sunucu, çerezler (bireysel istemci tarayıcılarının site ziyaretçilerini otomatik izlemesi için web sunucu tarafından üretilmiş işaretçiler) gibi diğer kullanışlı bilgileri de saklayabilmektedir (Batista ve Silva, 2001).

Her kullanıcı tanımlandıktan sonra her kullanıcı için tıklama akışı oturumlara bölünmelidir. Kullanıcının ne zaman web sitesini terk ettiği bilinemediğinden, kullanıcı oturumları oluşturulurken bir kesme noktası olarak genellikle zaman aşımı kullanılmaktadır (Joshi, Joshi ve Yesha, 2003).

Sonraki aşama ise örüntü keşif aşamasıdır. Bu aşamada kullanılan yöntemler ve algoritmalar istatistik, makine öğrenmesi ve veri tabanı gibi farklı alanlarda geliştirilmiştir. Web kullanım madenciliğinin bu aşamasında kullanılan üç yöntem vardır. Bunlar; birliktelik (hangi sayfalara birlikte erişildi), kümeleme (kullanıcı, işlem ve sayfa gruplarının bulunması) ve sıralı analizdir (web sayfalarına hangi sırada erişildi) (Batista ve Silva, 2001).

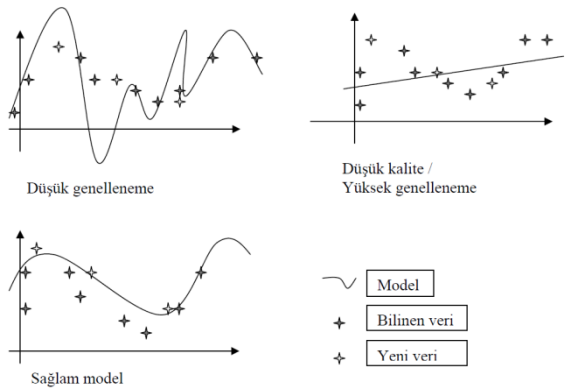
Örüntü analizi web kullanım madenciliğinin son aşamasıdır. Bu aşamada önceki aşamada bulunan ilgisiz örüntü ve kurallar elenmektedir. Görselleştirme teknikleri keşfedilmiş örüntüleri analiz etmede oldukça kullanışlıdır (Brusilovsky, Kobsa ve Nejdl, 2007).

3. Destek Vektör Makineleri

Temelleri ilk olarak Vladimir Vapnik ve Alexey Chervonenkis tarafından, hesaplanabilir öğrenme teorisinin önemli parçasını oluşturan ve öğrenmenin temel teorisi olarak bilinen Vapnik-Chervonenkis Teorisi kapsamında, 1960'lı yıllarda atılan destek vektör makineleri (DVM); 1992 yılında Vladimir Vapnik, Berhard Boser ve Isabelle Guyon tarafından sunulmuştur (Boser, Guyon ve Vapnik, 1992). Diğer sınıflandırma yöntemleri ile karşılaştırıldığında, eğitim süresi oldukça uzun olmasına rağmen, yüksek güvenirliği, ezbere öğrenmeye olan dayanıklılığı ve doğrusal olmayan sınıflandırmadaki başarı düzeyleri ile DVM tercih edilen bir yöntem olmuştur. DVM karar doğrusuna bağımlı olarak belirlenen destek noktaları (support vectors) arasındaki genişliği (margin) maksimize etmeyi amaçlayan danışmanlı bir öğrenme algoritmasıdır (Akpınar, 2014: 268).

Öğrenen makinenin genelleme kabiliyeti, öğrenen makineyle gerçeklenmiş fonksiyonlar setinin kapasitesine bağlıdır. Genelleme hatası tüm veri kümesi için gerçek risk ile eğitim örnek kütleleri için ampirik risk arasındaki fark olup örnek kütle sayısı arttıkça ve ağın kapasitesi azaldıkça azalmaktadır (Akaho, 1993: 493). Bu zayıf genelleme durumu aşırı uyumdan kaynaklanmaktadır. İstatistikte aşırı uyum çok fazla parametreye sahip istatistiksel modeli uygun hale getirmektedir. Garip ve yanlış bir model eldeki veri büyüklüğüyle karşılaştırıldığında yeterli karmaşıklığa sahipse mükemmel uyum gösterebilir. Buna aşırı uyum (overfitting) sorunu adı verilmektedir ve şu şekilde tanımlanmaktadır: Bir h hipotezi eğitim verisine, eğer eğitim verisi üzerinde h'nin h^''den daha küçük hataya sahip olduğu bir başka h^' hipotezi varsa, aşırı uyum sağlar. Ancak h^', test verisi üzerinde h'den daha küçük hataya sahiptir. Bu problem tüm öğrenme algoritmaları için geneldir (Wang ve Zhang, 2003).

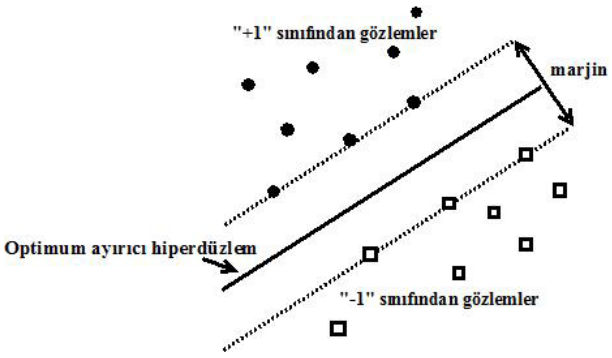
Aşırı uyum kavramı makine öğrenmesinde önemlidir. Eğitim süresince durumlara çıkarımsal ikileme dayanarak makinenin diğer örnekler için doğru çıktıyı öngörebilecek bir duruma ulaştığı varsayılmaktadır. Ancak makine hedef çıktıyla ilişkisi bulunmayan, eğitim verisinin nadir rassal özellikleri üzerine eğitimi düzenleyebilmektedir. Bu aşırı uyum sürecinde eğitim verisi üzerindeki performans artarken test verisi üzerindeki performans kötüleşmektedir (Şekil 1).



Şekil 1: Modelin Genellemesi (Tolun, 2008: 61)

3.1. Doğrusal Sınıflandırma (Marjlar ve Uzaklıklar)

Destek vektör makinelerinin ardında yatan düşünce, verileri optimal olarak iki sınıfa ayıran bir sınıflandırıcı oluşturulması problemidir (Clarke, Fokoué ve Zhang, 2009). İyi bir ayırıcı düzlem düşüncesi, veriden çok uzakta geniş bir marj (maksimal marj) sınıflandırıcı kavramında şekillenmektedir. Marj, iki veri bulutunu ayıran boş şerit genişliğini ifade etmektedir. Şekil 2'de iki sınıfın mükemmel bir şekilde ayrıldığı bir örnek verilmiştir. Bu veri seti için bile gözlemleri ayırabilen sonsuz sayıda hiperdüzlem bulunmaktadır. Ancak istenen, gelecekteki gözlemleri de doğru sınıflandıracak bir sınırın seçilmesidir.



Şekil 2: İdeal durum. İki veri bulutundaki noktalar arasındaki minimum (dik) uzaklık marjdır.

Marj, hiperdüzlem ve gözlemler arasındaki minimum (dik) uzaklık olarak tanımlanmaktadır. Amaç, en küçük (minimal) uzaklık kriterini sağlayan düzlemler arasındaki marjın maksimizasyonudur (minimum uzaklıkların maksimumunun bulunması problemi).

Şekil 3'teki kesikli çizgiler en yakın iki hiperdüzlemi göstermektedir. Veri noktaları en küçük (dik) uzaklığı gerçeklemektedir. Marj, dıştaki iki hiperdüzlem arasındaki uzaklıktır. Düz çizgi optimum ayırıcı hiperdüzlemi (geniş marj sınıflandırıcı) göstermektedir (Clarke, Fokoué ve Zhang, 2009).

Merkezi marjın konumunu biçimlemek için dört işlem bulunmaktadır:

1) a noktası ve ayırıcı hiperdüzlem arasındaki uzaklığı hesaplamak

2) verilen gözlemler setinin minimum uzaklığını sağlamak

3) iki paralel hiperdüzlem arasındaki uzaklığı hesaplamak

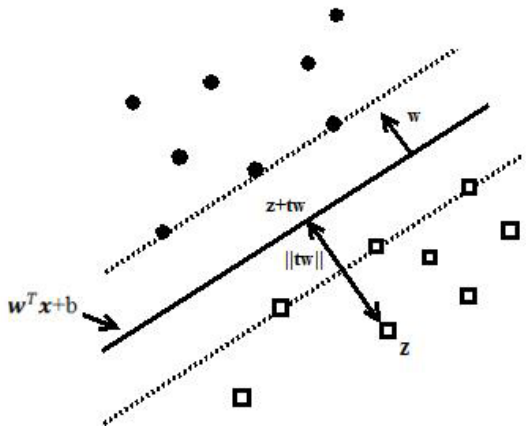
4) ayırıcı hiperdüzlemden eşit uzaklıktaki iki hiperdüzlemi sağlamak

Bu adımlar geometrik olarak sezgisel görünse de bazı tanımlamalar gerektirmektedir.

$w = \left(w_1, w_2, \dots, w_p\right)^T \in \mathbb{R}^p$ bir vektör katsayısı ve $b \in \mathbb{R}$ bir sabit olsun. h: $\mathbb{R}^p \to \mathbb{R}$ olmak üzere;

$$h(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}+\mathrm{b}$$

doğrusal fonksiyonu yazılır.



Şekil 3: z noktası w yönünde çizgiyle gösterilen hiperdüzlemden t birim uzaklıktadır.

Verilen bir $c \in \mathbb{R}$ sabiti için,

$$H_c(\boldsymbol{w}, b) = \{\boldsymbol{x}: h(\boldsymbol{x}) = c\}$$

bir $(p-1)$ boyutlu hiperdüzlemdir. Eğer $c = 0$ ise, $H_{c=0}(\boldsymbol{w}, b)$, $H(\boldsymbol{w}, b)$'dir.

Bir vektör hiperdüzleme paralel ise hiperdüzlemin yönlü vektörü, hiperdüzlemin tüm mümkün yönlü vektörlerine dik ise hiperdüzleme normal vektördür. Açıkça $\boldsymbol{w}$ vektörü herhangi c için $H_c(\boldsymbol{w}, b)$ 'dir. Aslında

$\forall \boldsymbol{x}_i, \boldsymbol{x}_j \epsilon H_c(\boldsymbol{w}, b)$ için $\boldsymbol{w}^T \boldsymbol{x}_i + b = c = \boldsymbol{w}' \boldsymbol{x}_j + b$ ve buradan $\boldsymbol{w}^T (\boldsymbol{x}_i - \boldsymbol{x}_j) = 0$ dır.

DVM sınıflandırıcı formülasyonu, $\mathbb{R}^p$'de bir nokta ile bir hiperdüzlem arasındaki dikey uzaklığın ifadesini gerektirmektedir.

Teorem: $H_c(\boldsymbol{w}, b)$ hiperdüzlemi ve $z \epsilon \mathbb{R}^p$ olmak üzere bir nokta arasındaki dikey uzunluk $d(z, H_c(\boldsymbol{w}, b))$ olmak üzere;

$$d(z, H_c(\boldsymbol{w}, b)) = \frac{|\boldsymbol{w}^T z + b - c|}{\|\boldsymbol{w}\|}$$

Teorem: $H_c(\boldsymbol{w}, b)$ ve $H_{c'}(\boldsymbol{w}, b)$ iki paralel hiperdüzlemi arasındaki dikey uzaklık;

$$d(H_c(\boldsymbol{w}, b), H_{c'}(\boldsymbol{w}, b)) = \frac{|c - c'|}{\|\boldsymbol{w}\|}$$

3.1. Doğrusal Olmayan Sınıflandırma

Doğrusal olmayan sınıflandırma destek vektör makineleriyle yumuşak marj (aylak değişken) ve kernel hilesi (kernel trick) yöntemi olmak üzere iki şekilde ele alınabilmektedir. Maksimal marj sınıflandırıcıyla ilgili temel sorun her zaman mükemmel şekilde tutarlı yani eğitim hatası olmayan bir hipotez üretmesidir. Marj kavramına bağımlılık, gürültünün her zaman var olabileceği gerçek verilerde kırılgan bir tahminleyene neden olabilmektedir. Yumuşak marjlar hatayı minimize ederek doğrusal olarak ayrılamayan sınıfları çözmeye yardımcı olmaktadırlar. Kernel hilesi yöntemi ise doğrusal olmayan DVM'ler için kullanılmaktadır ve girdi vektörlerini özellik uzayına haritalamaktadır.

3.1.1. Yumuşak Marj DVM:

Gürültünün var olduğu veya sınıfların üst üste bindiği durumlarla başa çıkabilmek için marjların yerleştirilmesinde; marj içerisindeki tüm veriler, ister ayırıcı doğrunun yanlış tarafında ister doğru tarafında olsun ihmal edilerek, hatayla işbirliği yapmak mümkündür. Doğrusal ayrılabilir olmayan durumda optimal hiperdüzlem geometrik marjı maksimize eden ve hata fonksiyonunu minimize eden bir hiperdüzlem olarak tanımlanmaktadır. Kısıtların kontrollü olarak ihlal edilmesine izin veren marj aylak değişkenlerinin ξ_i eklenmesi yoluyla $\mathrm{y_i}((\mathrm{x_i} * \psi) + \mathrm{b}) \geq 1$, $\mathrm{i} = 1,2, \ldots . \ell$ kısıtları $\mathrm{y_i}((\mathrm{x_i} * \psi) + \mathrm{b}) \geq 1 - \xi_i$, $\xi_i \geq 0$ olarak değişmektedir. Bu durumda hem marj maksimizasyonunu hem de hata minimizasyonunu birlikte ifade eden optimizasyon problemi (1) (Vapnik, 1998: 411) şu şekilde ifade edilmektedir:

$$Min\ \phi(\psi, \xi) = \tfrac{1}{2}(\psi * \psi) + C\left(\textstyle\sum_{i=1}^{\ell} \xi_i\right) \qquad (1)$$

Öyle ki,

$$\mathrm{y_i}((\mathrm{x_i} * \psi) + \mathrm{b}) \geq 1 - \xi_i,$$

$$\xi_i \geq 0, \mathrm{i} = 1,2, \ldots . \ell$$

Eğer $\xi_i > 1$ ise x_i veri noktası hiperdüzlem tarafından yanlış sınıflandırılmıştır. Ceza faktörü olarak bilinen C, eğitim hatasını minimize etmek ile marjı maksimize etmek arasındaki eşiği kontrol etmektedir ve kullanıcı tarafından belirlenmektedir.

Lagranj tekniği ile çözülen optimizasyon problemi (1) dual formda da ifade edilebilmektedir. Dual form problemin iç çarpımlar cinsinden ifade edilmesine fırsat tanıdığı için orijinal formülasyonu tercih edilmektedir.

$$Maks\ W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j (x_i * x_j)$$

Öyle ki,

$$0 \leq \alpha_i \leq C, i = 1, \ldots, \ell,$$

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0$$

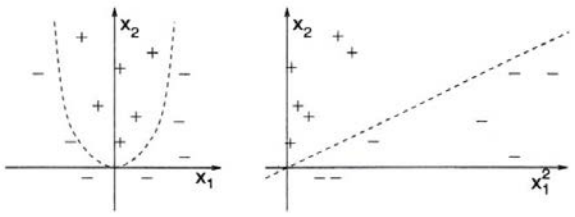
$\alpha_i > 0$ olan tüm eğitim örnekleri daha önce olduğu gibi destek vektörleri olarak adlandırılmaktadır. Destek vektörleri artık ya da marj üzerinde [$0 < \alpha_i < C$'ya sahip $(y_i f(x_i) = 1)$] (sınırlandırılmamış destek vektörleri) ya da marj alanının içerisinde uzanmaktadırlar [$\alpha_i = C$'ya sahip $y_i f(x_i) < 1$)] (sınırlandırılmış destek vektörleri).

3.1.2. Kernel Hilesi

DVM'ler doğrusal olmayan öğrenenlere kolaylıkla dönüştürülebilmektedir. Bu süreçte orijinal verilerden sınıflandırma özelliklerini çıkarmak için doğrusal olmayan bir haritalama kullanılmaktadır (Pöyhönen, 2004). Haritalama fonksiyonunda ana fikir, verinin daha üst boyutlarda bir uzaya dönüştürülmesidir (Akpınar, 2014: 277). Bir kernel (çekirdek) fonksiyonu genellikle veriyi özellik uzayına $\phi(\mathrm{x})$ fonksiyonu ile haritalayarak onun boyutsallığını arttırmaktadır. DVM daha sonra özellik uzayında maksimal marj doğrusal sınıflandırma kuralını öğrenmektedir. Sınıflandırma kuralı özellik uzayında doğrusal olsa da, orijinal girdi uzayında doğrusal değildir.

Şekil 4'te sol taraf (x_1, x_2) 'de doğrusal ayrılabilir olmayan eğitim kümeleri gösterilmektedir. Sağ tarafta ise aynı problem, (x_1^2, x_2) düzlemi üzerine yansıtılarak

doğrusal olmayan bir dönüşüm sonrasını göstermektedir. Bu yeni uzayda eğitim örnekleri doğrusal ayrılabilirdir.



Şekil 4: Doğrusal olmayan haritalama (Joachims, 2002).

Doğrusal olmayan haritalama ϕ; destek vektörleri ile özellik uzayındaki örüntü vektörü arasındaki iç çarpımları hesaplamak için kernel fonksiyonlarını kullanmaktadır:

$$K(x_i, x_j) \equiv (\emptyset(x_i)^T \emptyset(x_j))$$

Literatürde en sık kullanılan kernel tipleri ve parametreleri Tablo 1'de (bkz. Ekler) gösterilmiştir.

4. Analiz

E-ticaret sitelerinde ziyaretçiler her zaman satın alım kararı ile web sitesini ziyaret etmemekte; ürün kıyaslama, fiyat öğrenme gibi amaçlarla da sitelere erişmektedir. Bu durum satın alım sayısının ziyaret sayısından oldukça küçük olmasına, başka bir ifadeyle müşteri tepki oranının düşük kalmasına yol açmaktadır. Müşteri tepki oranının düşük olduğu dolandırıcılık tespiti, iptalden vazgeçirme, promosyon çalışmaları v.b. alanlar Müşteri tepki modellerini ortaya çıkartmıştır. E-ticaret alanında ise müşteri tepki oranı diğer alanlara göre çok daha düşüktür. Bu çalışmada kullanılan e-ticaret web sitesi verisi için bu oran yüzde birden daha küçüktür. Literatürde, müşteri tepki modelleri çerçevesinde çok sayıda uygulama olsa da e-ticaret alanında web günlük verileriyle yapılmış çalışmaya rastlanmamıştır. Makine öğrenmesi yöntemleri ise müşteri tepki modellerinde oldukça seyrek kullanılmakta, ancak bu çalışma kapsamında bir çalışmaya rastlanılmamıştır. Bu çalışmada, makine öğrenmesi yöntemlerinden DVM'nin web kullanım madenciliği kapsamında müşteri tepki modeli olarak kullanılabilirliği test edilmiştir. Tepki modellerinde sıkça kullanılan çok değişkenli güçlü bir istatistik yöntem olan lojistik regresyon ile de sonuçlar karşılaştırılmıştır. Çalışmada Srivastava v.d.'nin (2000) önerdiği web madenciliği analiz süreci takip edilmiştir.

4.1. Verinin Ön Analizi

Web kullanım madenciliğinde web sunucu günlükleri kullanılmaktadır. Web sunucu günlükleri ASCII biçiminde metin dosyası olarak sunucularda saklanmakta, içeriğinde siteye giriş yapan ziyaretçilerin IP adreslerini de barındırmaktadır. IP gizliliğini esas alan şirketler web sunucu günlüklerini paylaşmamayı tercih etmektedir. Uygulamada kullanılan e-ticaret sunucu günlük dosyaları, İstanbul'da faaliyet gösteren bir e-ticaret sitesine aittir. E-ticaret sitesinin fiziksel satış mekânı bulunmamakta, sadece tek bir ürün (farklı seçeneklerinin olduğu) satışı gerçekleştirilmektedir. E-ticaret sitesinin isteği üzerine çalışmada ismi saklı tutulmuş, analiz ekranlarında ziyaretçi IP bilgileri gizlenmiştir.

Uygulamada söz konusu e-ticaret sitesinin 13 Mayıs 2011 ile 1 Ağustos 2013 arası dönemi içeren 812 güne ait sunucu günlük dosyaları kullanılmıştır. Verinin özet istatistiklerin çıkarılması için "Nihuo Web Log Analyzer" programı kullanılmıştır. Program ücretli olmakla birlikte 30 günlük deneme süresi içerisinde özet istatistikler çıkarılmıştır.

Tablo 2'de verilen genel istatistiklere bakıldığında toplam istek sayısının 23.538.873 olduğu görülmektedir. Bu sayı web günlük dosyalarındaki toplam satır sayısıdır. Her satır, ziyaretçiler tarafından sunucuya yapılmış bir isteğe karşılık gelmektedir. Bu istekler metin, resim ya da video olabilmektedir. Ayrıca ziyaretçilerden gelen isteklerden farklı olarak yazılımlar aracılığıyla yapılan istekler de bulunmaktadır. Örümcek (robot, bot, spider, crawler vb. istekler uygulamada ortak isimle örümcek olarak adlandırılmıştır ancak aralarında farklılıklar bulunmaktadır) adını verdiğimiz bu ziyaretler, başta Google olmak üzere diğer arama motorlarının ya da web üzerinden veri toplamak isteyen kişilerin oluşturduğu yazılımlar aracılığıyla yapılmış insan kaynaklı olmayan isteklerdir.

Ceplenmiş istekler, istemci üzerinde kaydedilmiş isteklerdir. Bir tarayıcı daha önce istenmiş dosyanın kopyasını yedeklemiş ise sunucuya güncel olanı değil yedeği gönderir. Kaybedilen istekler ise hatayla sonuçlanan isteklerdir.

Sayfa görüntüleme, web sitesinin tek bir web sayfasına erişimini ifade etmektedir. İstek sayısı web sunucuda bulunan bir dosyaya resim, metin, v.b. olan erişime karşılık gelirken, sayfa görüntüleme sayısı belli bir zamanda erişilen sayfa sayısıdır. Örneğin bir web sayfası 5 resim içeriyorsa o sayfaya olan ziyaret 6 istek olarak günlük dosyasında kayda alınacaktır. Bir istek web sayfasına, 5 istek ise resimlere olacaktır. Bir ziyaretçi 10 sayfa dolaşmış ve her sayfa 10 resim içeriyorsa web sunucu 110 istek kaydedecektir. Sayfa görüntüleme sayısı ise 10 olacaktır. Ziyaret sayıları, sayfa görüntülemenin IP adreslerine göre gruplanmış hali (tekil kullanıcıların ziyaretleri) olarak ifade edilebilir.

Tablo 2'de toplam istek sayısının 605.239'u örümcek denilen makine üretimidir. Örümcekleri, insanların değil çoğunlukla arama motorlarının web sayfalarına indeksleme amaçlı tesadüfi yaptıkları ziyaretler olarak tanımlamak mümkündür. Sayfa görüntülemelerine bakıldığında 2 milyonun üzerinde sayfa görüntülendiği görülmektedir. Günde ortalama 2.724 sayfa görüntülenmiştir. Ziyaret başına ise ortalama 4,87 sayfa görüntülenmiştir.

Ziyaret istatistiklerine bakıldığında toplamda 453.304 ziyaret gerçekleşmiş olmasına karşın, bunlardan sadece 253.505 adetinin bireylere ait olduğu gözükmektedir. Bir ziyaretçinin ortalama 6 dakika 11 saniye söz konusu e-ticaret sitesinde zaman geçirdiği görülmektedir.

4.2. Veri Temizleme ve Kullanıcı Tanımlama

Veri temizleme aşamasında bütün web günlük dosyaları phpmyadmin aracılığıyla veri tabanı oluşturularak bir araya toplanmıştır. SQL (Structured Query Language – Yapılandırılmış Sorgulama Dili) ile veri temizleme ve düzenleme işlemleri gerçekleştirilmiştir. Veri tabanından elde edilen düzenlenmiş veri .csv uzantılı olarak aktarılarak analizde kullanılabilecek hale getirilmiştir.

Nihuo Web Log Analyzer programından veri temizleme aşamasında da yararlanılmış, analizde istenmeyen verilerin elenmesi ve yazılımların yapmış olduğu sahte ziyaretler (robot) belirlenmiştir.

Çalışmada yararlanılan verinin ait olduğu e-ticaret sitesi, günlük dosyalarını Windows IIS biçiminde saklamaktadır. Bu günlük dosyaların ilk haline ait bir kesit (1 isteğe ait) aşağıda gösterilmiştir. IP numaraları kişisel güvenlik nedeniyle saklanmıştır (IP numaraları xx.xx.xxx.xxx ile gösterilmiştir).

2011-05-28 00:17:28 W3SVC1 xx.xx.xxx.xxx GET /DXR.axd r=2_15-okUJ2 80 - xx.xx.xxx.xxx Mozilla/4.0+(compatible;+MSIE+8.0;+Windows+NT+5.1;+Trident/4.0;+GTB6.5;+.NET+CLR+1.1.4322;+.NET+CLR+2.0.50727;+InfoPath.2;+OfficeLiveConnector.1.3;+OfficeLivePatch.1.3;+.NET+CLR+3.0.04506.30;+.NET+CLR+3.0.4506.2152;+.NET+CLR+3.5.30729) 200 0 0

Öncelikle bu metin dosyaları "mysql" yardımıyla sütunlarına ayrılarak veri tabanına aktarılmıştır. İlk etapta 9 sütuna ayrılmıştır. Bu sütunlar Windows IIS biçimli günlük dosyasının alanlarını temsil etmektedir. Her sütun bir niteliğe karşılık gelmektedir. Bu özellikler id, tarih (date), zaman (time), phptime, sunucu bilgisi (server), yöntem (method), dosya adı (filename), parametre ve istemci IP (clientip)'sidir.

Ardından analiz için gerekli olmayan resim, video, animasyon gibi dosyalar ayıklanmıştır. Filtrelenen uzantılar .swf, .jpeg, .gif gibi medya dosyalarıdır. Bu dosyaların tespiti için de yine Nihou web analyzer programından yararlanılmıştır.

Eleme işlemleri neticesinde 23.535.873 satır 1.872.097 satıra indirilmiştir. Veri temizleme işlemi sonlandıktan sonra web kullanım madenciliğinde önemli olan kullanıcı oturum tanımlama sürecine geçilmiştir. Kullanıcı tanımlama veri tabanına aktarım ile IP adreslerine göre bir sütunda tanımlanarak gerçekleştirilmiştir.

Oturum tanımlama, kullanıcı hareketleri sırasında toplanmış IP adreslerinin tüm sayfa görüntülenme kayıtlarını gruplamasından oluşmaktadır. Oturum tanımlama işleminin yapılmasının nedenini şu şekilde açıklamak mümkündür. Bir kullanıcı 24 saat içinde bir web sitesini iki kez ziyaret etmiş olsun. Ziyaretler 6 saat aralık ile yapılmış ise bu 24 saatlik periyot için kullanıcı tanımlama yöntemleri erken uygulanırsa, iki ziyaret beraber sıralanacak ve bu kullanıcıyla tanımlanacaktır. Ancak bu iki ziyaret arasında fark olması gerekir. Çalışmada ziyaretçi hareketsiz kalma süresi (ikinci oturum başlatma süresi) 60 dakika olarak belirlenmiştir. Oturumlar tanımlanırken kullanıcının siteye giriş yaptığı gün, kaç sayfa gezdiği, sitede ne kadar süre harcadığı, satın alma yapıp yapmadığı bilgileri hesaplanmıştır. Bu işlem için yine SQL'den ve ilave olarak PHP dilinden faydalanılmıştır. Oturum tanımlama ile 1.872.097 olan satır sayısı 91.095 oturuma dönüşmüştür.

Satışa dönüşen ziyaretler de oturum tanımlama ile belirlenmiştir. 91.095 oturumdan 366'sı satış ile sonuçlanmıştır. Siteye giriş yapılan saatler 4 kategori olarak 4 saat dilimine dönüştürülmüştür. Bu dört kategori Sabah (06:00-11:59 arası), Öğle (12:00-17:59 arası), Akşam (18:00-23:59 arası) ve Gece (00:00-05:59 arası) olarak tanımlanmıştır.

4.3. Veri Analizi

Destek vektör makineleri ve lojistik regresyon analizi için RapidMiner Studio 6 kullanılmıştır. RapidMiner, Java diliyle yazılmış, GNU altında açık kaynak kodlu bir veri madenciliği yazılım paketi olup, Excel, Access, Oracle, MySQL, SPSS gibi farklı veri kaynaklarına erişim sağlayan veri yükleme, dönüştürme, modelleme ve görselleştirmede güçlü bir araçtır. Veri temizlemeden sonraki adımlarda sınıflandırma yöntemleri RapidMiner aracılığıyla gerçekleştirilmiştir.

Analizler İstanbul Üniversitesi Bilimsel Araştırmalar Proje Birimi (BAP) projesi ile tedarik edilen 16 GB RAM'e sahip Intel Xeon CPU 3.50 GHz işlemcili workstation ile yapılmıştır. Analizde sınıflandırma yöntemi olarak destek vektör makineleri kullanılmıştır. Veri temizleme ve oturum tanımlama işlemleri sonucu veriden elde edilen nitelikler Tablo 3'de verilmiştir. Bunlar; kullanıcının sitede gezdiği sayfa sayısı,

ziyaretçinin sitede harcadığı toplam süre, oturumuna bağlı olarak sitede kaldığı saat dilimi [Sabah (06:00-11:59 arası), Öğle (12:00-17:59 arası), Akşam (18:00-23:59 arası), Gece (00:00-05:59 arası)], ziyaretçinin hangi ay siteye giriş yaptığı (Ocak, Şubat, Mart, Nisan, Mayıs, Haziran, Temmuz, Ağustos, Eylül, Ekim, Kasım, Aralık), siteye giriş yapılan gün (Pazartesi, Salı, Çarşamba, Perşembe, Cuma, Cumartesi, Pazar), hangi ülkeden istek yaptığı (Türkiye ve diğer ülkeler olarak iki kategoride tanımlanmıştır) ve son olarak da ziyaretçinin ürün satın alıp almadığına (satın alım için Evet, satın almama için Hayır) ait niteliklerdir. Satın alma niteliği kategorik olarak tanımlanmasına karşın aynı zamanda niceldir. Web sitesi özellikli ürün sattığı için her satın alma bir ürüne denk gelmektedir.

Klasik destek vektör makineleri sınıflandırma modeli şu şekildedir:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

x_i eğitim vektörleri, $\emptyset$ kernel fonksiyonu tarafından daha yüksek boyutlu bir uzaya haritalanmaktadır. Kernel fonksiyonu, $K(x_i, x_j) \equiv (\emptyset(x_i)^T \emptyset(x_j))$ olup, analizde sigmoid fonksiyonu kullanılmıştır:

$$K(x_i, x_j) = \tanh(\gamma x_i^T x_j + c)$$

DVM sınıflandırıcı ise,

$$sign(\sum_{i}^{l} \alpha_i y_i K(x, x_i) + b)$$

olarak tanımlanmaktadır.

91.095 oturuma karşılık 366 satış gerçekleşmiştir. Tepki oranı %1'in altında oldukça düşük bir yüzdeye sahiptir. Bu durum veri dengesizliği problemini ortaya çıkarmaktadır. Veri dengesizliğini gidermek amacıyla yeniden örnekleme yönteminden yararlanılmıştır. Yeniden örnekleme yöntemlerinden örnek küçültme kullanılarak çoğunluk sınıfından (satın almama durumu), azınlık sınıfının (satın alma durumu) 1, 2, 3 ve 4 katına denk gelecek şekilde tesadüfi olarak alt örnek veri setleri oluşturulmuştur. Bu şekilde oluşturulan dört modelde sınıflandırıcı performansları karşılaştırılmıştır. Bu dört modele ilave olarak bu kez azınlık sınıfından örnekler azaltılarak, azınlık sınıfın 1, 2, 3 ve 4 katına eşit olacak şekilde çoğunluk sınıfından örnekler çekilerek de analizler tekrarlanmıştır. Toplamda 20 model ile analizler gerçekleştirilmiştir.

Destek vektör makineleri ve lojistik regresyon için model oluşturulurken ilk önce veri seti satın alma (pozitif sınıf, 366 oturum) ve satın almanın gerçekleşmediği (negatif sınıf, 90.729 oturum) durumlara göre ikiye ayrılmıştır. Negatif sınıftan pozitif sınıfın 1, 2, 3 ve 4 katına denk gelecek şekilde tesadüfi olarak örnek seçilmiştir. Negatif sınıftan örnek seçilirken frekans aralığı, tarihlere göre sıralandırılmış, yıllara göre 3 gruba ayrılmıştır. Üç yıllık veriye sahip olmamızın yanı sıra üç tam yıl olmadığından (verinin ait olduğu dönem Mayıs 2011 ile Ağustos 2013 arasıdır) yılların sahip oldukları haftalara göre bir gruplamaya gidilmiştir. Aynı şekilde satın alımın gerçekleştiği pozitif sınıf da üç gruba ayrılarak örnekleme oranları hesaplanmıştır. Tablo 4'te negatif sınıfa ve Tablo 5'te pozitif sınıfa ait oluşturulan gruplar ve her model için çekilen tesadüfi örnek sayısı verilmiştir.

Tablo 4'te 1. Grup örnekler 2011 yılına ait 30.074 oturum arasından, 2. Grup örnekler 2012 yılına ait 42.251 oturum arasından ve 3. Grup örnekler 2013 yılına ait 18.404 oturum arasından ilgili oranlara göre çekilmiştir. Aynı şekilde pozitif örnekler de Tablo 5'te gösterildiği üzere satın alımların gerçekleştirildiği oturumlar göz önüne alınarak oluşturulmuştur. 1. Grup örnekler 2011 yılına ait 103 oturumu, 2. Grup 2012 yılına ait 188 oturumu ve son olarak 3. Grup da 2013 yılındaki 75 oturumu tanımlamaktadır.

Oluşturulan her gruptan tablolardaki oranlarda tesadüfi örnekler çekilerek analiz aşamasına geçilmiştir.

Veri setlerinden nitelik seçimi ve normalizasyon işlemleri gerçekleştirildikten sonra verinin %80'i eğitim, %20'si test seti olarak ayrılmıştır.

Doğrusal, polinominal, radyal tabanlı ve sigmoid kernel tipleri denenmiş, en iyi sonucu sigmoid kernel vermiştir. Parametre değerleri C (hata teriminin ceza parametresi (C>0)) ve gamma (γ) için optimum değerler RapidMiner'da bulunan evrimsel parametre optimizasyon tekniği kullanılarak belirlenmiştir. Aşırı uyumdan kaçınmak, tesadüfiliği artırmak amacıyla çapraz değerleme yöntemi kullanılmıştır. Eğitim seti 10 eşit alt sete bölünmüş ve setlerden 9 adedi eğitim, 1 adeti test seti olarak kullanılmıştır. Böylelikle 10 farklı eğitim ve test seti için parametre optimizasyon tekniği uygulanmıştır. Elde edilen parametreler modelin parametreleri olarak DVM eğitiminde kullanılmıştır. Nihai olarak da modeller test verisine uygulanarak sınıflandırıcı performansları elde edilmiştir. Çoğunluk ve azınlık sınıfın farklı oranlara göre oluşturulan modeller için hem DVM hem de lojistik regresyon analiziyle elde edilen sınıflandırma sonuçları karşılaştırılmıştır.

Destek vektör makineleri ve lojistik regresyon ile cevap oranı 1, 2, 3 ve 4 kat olacak şekilde modeller

oluşturulmuştur. Tablo 6'da destek vektör makineleri (DVM) ve lojistik regresyon analizi (LR)'nin doğruluk, kesinlik, duyarlılık ve belirlilik, kappa, F-skor değerleri verilmiştir.

Pozitif sınıfların doğru sınıflandırma yüzdesi, tepki oranı yüzde 50 olan modellerde en yüksek seviyeye ulaşmıştır. Negatif örnek sayısı arttıkça pozitif sınıf tahminleri düşmektedir. Örnek azaltmayla yapılan analizler (Model 1a, 2a, 3a, 4a) sonrasında pozitif sınıftan da örnek azaltılarak tekrarlanmıştır. Böylelikle tepki oranlarına göre değişimler gözlenmiştir. Tablo 6'da özet olarak verildiği üzere hem lojistik regresyon hem de destek vektör makineleri için en iyi pozitif sınıf tahmini negatif örnek ile pozitif örneğin eşdeğer olduğu modellerde sağlanmış, negatif örnek sayısı artıkça (tepki oranı azaldıkça) pozitif sınıf tahmin gücü azalmıştır.

Düzenlenen verilerden, DVM ve LR ile kurulan modellerin hem eğitimi hem testi için rassal olarak örnek veri setleri çekilmiştir ve bu işlem her iki yöntem için de 5'er kez yinelenerek farklı rassal veri setleri üzerinde denemeler yapılmıştır. Analizler yapıldıktan sonra ise 6 adet farklı performans ölçüm kriterine göre model değerlendirmeleri elde edilmiş ve rassal veri setleri için bulunan her bir performans ölçümünün ortalama değeri hesaplanmıştır.

Sonuçlara bakıldığında (Tablo 6) farklı performans ölçümleri için DVM ve LR modellerinin üstünlüklerinin de farklılık gösterdiği görülmektedir. Örneğin %50 tepki oranı için sonuçlara bakıldığında doğruluk, F skoru ve duyarlılık ölçümlerine göre DVM; kappa, belirlilik ve kesinlik ölçümlerine göre ise LR üstünlük göstermektedir. % 33 tepki oranında tüm performans ölçümlerinde LR üstünlük sağlarken, %25 oranında belirlilik haricinde diğer tüm performans ölçümlerinden DVM üstünlük sağlamıştır. Son olarak %20 tepki oranında ise kappa ve belirlilik ölçümlerinde DVM; doğruluk, F skoru, duyarlılık ve kesinlik ölçümlerinde ise LR üstünlük sağlamıştır.

Dört farklı tepki oranına göre (%50, %33, %25, %20) her bir performans ölçümü için elde edilen ortalama değerlerin de ortalaması alındığında ise, hem DVM hem de LR için en iyi sonucun yüzde 50 tepki oranı ile oluşturulan model olduğu ve DVM ile LR performans ölçümlerinin (0,818) birbirine eşit olduğu görülmektedir. Bir başka ifadeyle, söz konusu ikili sınıflandırma görevi için her iki sınıfa ait örnek sayılarının eşit olduğu durumda hem DVM hem de LR ile kurulan modellerde en iyi sınıflandırma sonuçlarına ulaşılmıştır.

Yukarıda özetlenen sonuçlar ışığında, %50 tepki oranı ile kurulan modelin analiz için tercih edilmesi gerektiği açıktır. Buna göre DVM ile %50 tepki oranı için kurulan modelde tüm rassal veri setleri için vektör ağırlıkları incelendiğinde "Türkiye", "sabah", "haziran" ve "ocak" niteliklerinin pozitif sınıfların (satın alım olması) belirlenmesinde; "Diğer", "kasım" ve "sayfa sayısı" niteliklerinin ise negatif sınıfların (satın alım olmaması) belirlenmesinde etkili olduğu görülmektedir.

LR için de en iyi model olan % 50 tepki oranı için rassal veri setlerinden en iyi sonucu veren 1e modelidir ve pozitif sınıfları F=%90 ve negatif sınıfları %100 kesinlikle doğru tahmin etmiştir. Modelde kalan 0,05 anlamlılık seviyesindeki nitelikler ile açıklayıcılık değeri 0,677 olan LR modeli; "= 0.71 + 0,29 * Türkiye - 0,33 * Sabah - 0,4 * Temmuz + 0,24 * Perşembe + 2,98 * Sayfa Sayısı" şeklinde ifade edilebilmektedir.

5. Sonuç

Bu çalışmada web kullanım madenciliği ile şirketlerin sunucularında tutulan web günlük verileri kullanılmıştır. Bu veriler metin dosyaları halindedir ve gereksiz, yararlı olmayan birçok veriye sahiptir. Bu gereksiz veriler kirli veri olarak adlandırılmaktadır.

Günlük verilerinden ziyaretçinin ülke bilgisi (IP adresinden çıkarılmıştır), site içerisinde dolaşılan sayfa sayısı, sitede harcanan süre, sitenin ziyaret edildiği gün, ay ve saat dilimi ve site ziyaretinde satın alım yapılıp yapılmadığına ilişkin nitelikleri içermektedir.

Ziyaretçiler e-ticaret sitelerine her zaman kesin satın alma isteğiyle gelmemektedirler. Potansiyel müşteriler ürün kıyaslama, fiyat öğrenme gibi nedenlerle de web sitesini ziyaret edebilmektedir. Bu nedenle web sitesi ziyaretlerinde sayfa / ürün görüntüleme sayılarıyla, ürün satışları arasında büyük farklar bulunmaktadır. Bu durum MİY odaklı alanlarda, gerçek uygulamalarda kullanılan veri setlerinde sıkça gözükmektedir. Tepki oranı düşüklüğü olarak da adlandırılan bu dengesiz veri durumu, sınıflandırma yöntemlerinde doğruluk oranlarının gerçekçi olmayacak şekilde artmasına neden olmaktadır.

Çalışmada kullanılan e-ticaret sitesine ait web günlük verilerinde satış cevap oranı yüzde birden daha azdır. Bu oran literatürde karşılaşılan oranlardan da düşüktür. Çözüm olarak satın alımların (pozitif sınıf) bir, iki, üç ve dört katı oranlarında negatif sınıftan tesadüfi örnekleme yöntemiyle negatif sınıf oluşturularak veri dengesizliği giderilmeye çalışılmıştır. Genelleme yapmak için pozitif örnekten azaltmalar yapılarak model sayısı arttırılmıştır.

Destek vektör makineleri ile elde edilen sınıflandırma sonuçları, pazarlama ve MİY alanında sıklıkla kullanılan lojistik regresyon analizi sonuçları ile karşılaştırılmıştır.

Düzenlenen verilerden, DVM ve LR ile kurulan modellerin hem eğitimi hem testi için rassal olarak örnek veri setleri çekilmiştir ve bu işlem her iki yöntem için de

5'er kez yinelenerek farklı rassal veri setleri üzerinde denemeler yapılmıştır. Analizler yapıldıktan sonra ise 6 adet farklı performans ölçüm kriterine göre model değerlendirmeleri elde edilmiş ve rassal veri setleri için bulunan her bir performans ölçümünün ortalama değeri hesaplanmıştır.

Sonuçlara bakıldığında (Tablo 6) farklı performans ölçümleri için DVM ve LR modellerinin üstünlüklerinin de farklılık gösterdiği görülmektedir. Dört farklı tepki oranına göre (%50, %33, %25, %20) her bir performans ölçümü için elde edilen ortalama değerlerin de ortalaması alındığında ise, hem DVM hem de LR için en iyi sonucun yüzde 50 tepki oranı ile oluşturulan model olduğu ve DVM ile LR performans ölçümlerinin (0,818) birbirine eşit olduğu görülmektedir. Bir başka ifadeyle, söz konusu ikili sınıflandırma görevi için her iki sınıfa ait örnek sayılarının eşit olduğu durumda hem DVM hem de LR ile kurulan modellerde en iyi sınıflandırma sonuçlarına ulaşılmıştır.

DVM için çeşitli kernel tipleri denenmiş ve test verisi üzerinde denemelerde en iyi sınıflama sonuçlarına sigmoid kernel ile ulaşılmıştır. LR'de de lojistik fonksiyon (sigmoid fonksiyon) kullanılmaktadır. Genel olarak tüm modellerde sonuçların yakın çıkmasının (özellikle tepki oranının yüzde 50 olduğu modelde aynı olması) nedeni olarak sigmoid fonksiyonun kullanılmış olması göz önünde bulundurulması gereken bir nedendir.

Yukarıda özetlenen sonuçlar ışığında, %50 tepki oranı ile kurulan modelin analiz için tercih edilmesi gerektiği açıktır. Buna göre DVM ile %50 tepki oranı için kurulan modelde tüm rassal veri setleri için vektör ağırlıkları incelendiğinde "Türkiye", "sabah", "haziran" ve "ocak" niteliklerinin pozitif sınıfların (satın alım olması) belirlenmesinde; "Diğer", "kasım" ve "sayfa sayısı" niteliklerinin ise negatif sınıfların (satın alım olmaması) belirlenmesinde etkili olduğu görülmektedir.

Veri ön analizlerinde, satın alma davranışında ziyaret edilen sayfa sayısının ve ziyaret süresinin baskın çıktığı görülmüştür. Modellerin çoğunda sayfa sayısı, süresi ve ülke = Türkiye, satın alım tahminini etkileyen nitelikler olarak öne çıkmaktadır. DVM sonuçları veri ön analizi sonucunda yapılan çıkarımlarla paralellik göstermektedir.

Çalışmada web günlük verilerinden elde edilen niteliklere göre satın alma davranışı elde edilmeye çalışılmıştır. Bir ziyaretçiyi satın almaya teşvik eden dış etkenler de bulunmaktadır. Kullanıcıların demografik bilgileri, meslekleri, aylık gelirleri gibi nitelikleri bu dış etkenlerden bazılarıdır ve uygulamada yer almamaktadır. Bu durum analizin önemli kısıtlarından biridir. Çalışmada veri kaynağı olarak sadece web günlük verileri kullanılabilmiştir. Kullanıcı IP'leri gizli olduğundan web günlük verisi elde etme aşamasında da zorluklar yaşanmıştır. Şirket ismi saklanarak ve IP numaraları gizlenerek analizler gerçekleştirilmiştir. Destek vektör makinelerinin daha etkin kullanımı için e-ticaret sitelerinin ilave müşteri bilgilerini saklamaları önerilmektedir.

Çalışmanın diğer bir kısıtı çalışmaya konu olan e-ticaret sitesinin sadece tek bir ürünün (özellikli ürün) değişik varyasyonlarını satmasıdır. Bu nedenle sayfalar arası ilişki kurulamamıştır. Bir ziyaretçi bir oturumda en fazla bir ürün satın almaktadır ve ücret ya da ürün sayısı da ulaşılamayan niteliklerdir. Satın alma durumunun nadir olduğu (%1'in altında) veri seti uygulamalarında veri seti küçültülerek alt veri setleri üzerinde analizler gerçekleştirilmiş, demografik verilerin ve ürün çeşitliliğinin olduğu veri setleri üzerinde yapılması sonraki çalışmalara bırakılmıştır.

Kaynakça

Agosti, M. ve Di Nunzio, G. (2007). Web Log Mining: A Study of User sessions. Proceedings of the 10th DELOS Thematic Workshop on Personalized Access, Profile Management, and Context Awareness in Digital Libraries (PersDL).

Akaho, S. (1993). VC Dimension Theory for a Learning System with Forgetting, Proceedings of 1993 International Joint Conference on Neural Networks, Tokyo, 1(25-29), 493-496.

Akpınar, H. (2014). DATA Veri Madenciliği Veri Analizi, Papatya Yayıncılık.

Batista, P. ve Silva, M. (2001). Mining Web Access Logs of an On-line Newspaper. The Proceedings of 12th International Meeting of the euro working group on decision support systems.

Bose, I. ve Chen, X. (2009). Quantitative models for direct marketing: A review from systems perspective. European Journal of Operational Research, 195(1), 1-16

Brusilovsky, P., Kobsa, A. ve Nejdl, W. (2007). The Adaptive Web, Springer.

Bucklin, R. (2008). Marketing Models for Electronic Commerce. Handbook of Marketing Decision Models, s. 327.

Burez, J. ve Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36, 4626-4636.

Chawla, N., Bowyer, K., Hall, L. ve Kegelmeyer, W.(2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16(1), 321-357.

Chitraa, V. ve Davamani, A. (2010). A Survey on Preprocessing Methods for Web Usage Data.

International Journal of Computer Science and Information Security, 7(3).

Clarke, B., Fokoué, E. ve Zhang, H. (2009). Principles and Theory for Data Mining and Machine Learning. New York: Springer Science+Business Media.

Cooley, R., Mobasher, B. ve Srivastava, J. (1997). Web Mining: Information and Pattern Discovery on the World Wide Web. Proceedings 9th IEEE International Conference on Tools with Artificial Intelligence, 558-567.

Drummond, C., Holte, R. (2003). C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling beats Over-Sampling. Workshop on Learning from Imbalanced Datasets II, ICML. Washington DC.

Etzioni, E. (1996). The World Wide Web: Quagmire or Gold Mine. Communication of the ACM, 39(11), 65-68.

He, H., Garcia, E. (2009). Learning from Imbalanced Data. Ieee Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

Hughes, A. (2005). Strategic Database Marketing: The Masterplan for Starting and Managing a Profitable, Customer-Based Marketing Program, McGraw-Hill.

Japkowicz, N. (2000). The Class Imbalance Problem: Significance and Strategies. In Proceedings of the 2000 International Conference on Artificial Intelligence: Special Track on Inductive Learning, 1.

Japkowicz, N. (2001). Concept learning in the presence of between class and within-class imbalances. In Proceedings of the Fourteenth Conference of the Canadian Society for Computational Studies of Intelligence. 67-77.

Jiawei, H., Kamber, M. ve Pei, J. (2011). Data Mining Concepts and Techniques, Morgan Kaufmann.

Joshi, K., Joshi, A. ve Yesha, Y. (2003). On Using a Warehouse to Analyze Web Logs. Distrubuted and Parallel Databases.

Khoshgoftaar, T., Van Hulse, J. ve Napolitano, A. (2011). Comparing Boosting and Bagging Techniques With Noisy and Imbalanced Data. Ieee Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, 41(3), 552-568.

Kim, G., Chae, B.ve Olson, D. (2013). A support vector machine (SVM) approach to imbalanced datasets of customer responses: comparison with other customer response models. Service Business, 7, 167-182.

Kohavi, R., Mason, L., Parekh, R. ve Zheng, Z. (2004). Lessons and Challenges from mining retail e-commerce data. Machine Learning, 57(1), 83-113.

Kosala, R. ve Blockeel, H. (2000). Web Mining Research: A Survey. SIGKDD Explorations: Newsletter of the Special Interest Group on *Knowledge Discovery & Data Mining*, 1-15.

Ling, C. ve Li, C. (1998). Data Mining for Direct Marketing: Problems and Solutions. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), 73-79.

Liu, B. (2007). Web Data Mining Exploring Hyperlinks, Contents, and Usage Data. Springer 1st ed.

Ngai, E., Xiu, L. ve Chau, D. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications, 36(2), 2592-2602.

McCarty, J. ve Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. Journal of Business Research, 60(6), 656-662.

Pöyhönen, S. (2004). Support Vector Machine Based Classification in Condition Monitoring of Induction Motors, Helsinki University of Technology Control Engineering Laboratory, Finland.

Provost, F. ve Fawcett, T. (2001). Robust Classification for Imprecise Environments. Machine Learning, 42(3), 203-231.

Srivastava, J., Cooley, R., Deshpande, M. ve Tan, P. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, 12-23.

Tolun S. (2008). Destek Vektör Makineleri: Banka Başarısızlığının Tahmini Üzerine Bir Uygulama, Doktora Tezi, İ.Ü. S.B.E.

Vapnik, V. (1998). Statistical Learning Theory, Wiley.

Viaene, S., Baesens, B., Van Gestel, T., Suykens, J., Van den Poel, D., Vantihienen, J. ve Dedene, G. (2001). Knowledge Discovery in a Direct Marketing Case using Least Squares Support Vector Machines. International Journal of Intelligent Systems, 16(9), 1023-1036.

Wang, X. ve Zhang, Y. (2003). Statistical Learning Theory and State of Art in DVM, Proceedings of the Second IEEE International Conference on Cognitive Informatics, IEEE, 55-59.

Weiss, G. (2004). Mining with rarity: a unifying framework. ACM SIGKDD Explorations Newsletter - Special issue on learning from imbalanced datasets, 6(1), 7-19.

Xu, G., Zhang, Y. ve Li, L. (2011). Web Mining and Social Networking Techniques and Applications. Springer Science+Business Media.

EKLER

Tablo 1: Kernel tipleri ve parametreleri

Kernel Tipi	Eşitlik	Parametre
Doğrusal	$x_i^T x_j$	-
Polinominal	$\left(\gamma(x_i^T x_j + c)\right)^p$	γ, c, p
Gaussian (RBF)	$exp\left(-\|x_i - x_j\|^2/\sigma^2\right)$	σ
Sigmoid (MLP)	$tanh(\gamma x_i^T x_j + c)$	γ, c

Tablo 2: İstek Sayısı, Ziyaret, Sayfa İstatistikleri

İstek Sayıları	
Toplam İstek Sayısı	23.535.873
Normal İstek Sayısı	22.930.634
Örümcek İstek Sayısı	605.239
Gün Başına Düşen Ortalama İstek Sayısı	29.020
Ziyaret Başına Düşen Ortalama İstek Sayısı	51,92
Ceplenmiş İstekler	1.261.170
Kaybedilen İstekler	310.900
Sayfa Görüntülemeleri	
Toplam Sayfa Görüntülemesi	2.209.288
Gün Başına Düşen Ortalama Sayfa Görüntülemesi	2.724
Ziyaret Başına Düşen Ortalama Sayfa Görüntülemesi	4,87
Tekil Sayfa Görüntülemesi	1.428.643
Ziyaretler	
Toplam Ziyaret Sayısı	453.304
Normal Ziyaretler	253.505
Örümcek Ziyaretler	199.799
Gün Başına Düşen Ortalama Ziyaret	558
Toplam Ziyaretçi Kalış Süresi (dakika)	46.799:57:02
Ortalama Ziyaretçi Kalış Süresi (dakika)	06:11

Tablo 3: Analizde kullanılan nitelikler

Nitelikler	**Özellikleri**
Sayfa Sayısı (sayfasayisi)	Nicel
Süre (sure)	Nicel
Saat Dilimi (saatdilimi)	4 kategori: Sabah(06:00-11:59) Öğle (12:00-17:59) Akşam(18:00-23:59) Gece (00:00-05:59)
Ay (ay)	12 kategori
Gün (gun)	7 kategori
Ülke (ulke)	2 kategori: Türkiye Diğer
Satın Alma Durumu (satinalim)	Kategorik Satın alma için EVET Satın almama için HAYIR

Tablo 4: Negatif Sınıfa Ait Örnek Oranları

		Gruplar (Oranlar)		
Model	Negatif Örnek Sayısı	1 (1-30.074)	2 (30.075-72.325)	3 (72.326-90.729)
Model 1a	366	0,0035	0,0039	0,0052
Model 2a	732	0,0070	0,0078	0,0100
Model 3a	1098	0,0140	0,0120	0,0150
Model 4a	1464	0,0140	0,0160	0,0200
Model 4b	1200	0,0120	0,0130	0,0170
Model 3b	900	0,0086	0,0096	0,0130
Model 4c	800	0,0076	0,0086	0,0110
Model 2b-3c	600	0,0058	0,0064	0,0080
Model 2c-4d	400	0,0038	0,0043	0,0050
Model 1b-3d	300	0,0029	0,0032	0,0040
Model 1c-2d-4e	200	0,0019	0,0021	0,0030
Model 3e	150	0,0014	0,0016	0,0020
Model 1d-2e	100	0,0010	0,0010	0,0010
Model 1e	50	0,0005	0,0005	0,0007

Tablo 5: Pozitif Sınıfa Ait Örnek Oranları

		Gruplar		
Model	Pozitif Örnek Sayısı	1 (1-103)	2 (104-291)	3 (292-366)
Model 1a-2a-3a-4a	366	1	1	1
Model 1b-2b-3b-4b	300	0,83	0,72	1
Model 1c-2c-3c-4c	200	0,56	0,48	0,7
Model 1d-2d-3d-4d	100	0,28	0,24	0,35
Model 1e-2e-3e-4e	50	0,14	0,12	0,17

Tablo 6: DVM ve LR Modellerinin Karşılaştırılmalı Performansı

		Doğruluk		Kappa		F		Duyarlılık		Belirlilik		Kesinlik		Ort. Performans	
Tepki Oranı	Model	DVM	LR	DVM	LR	DVM	LR	DVM	LR	DVM	LR	DVM	LR	DVM	LR
	1a	0,869	0,857	0,763	0,712	0,867	0,849	0,849	0,808	0,889	0,905	0,886	0,894		
	1b	0,862	0,783	0,672	0,664	0,868	0,780	0,875	0,754	0,848	0,814	0,862	0,807		
50%	1c	0,845	0,747	0,683	0,701	0,854	0,756	0,864	0,750	0,825	0,744	0,844	0,762		
	1d	0,833	0,878	0,701	0,699	0,844	0,894	0,864	0,955	0,800	0,790	0,826	0,840		
	1e	0,810	0,905	0,578	0,773	0,818	0,900	0,818	0,818	0,800	1,000	0,818	1,000		
Ortalama		0,844	0,834	0,679	0,710	0,850	0,836	0,854	0,817	0,832	0,851	0,847	0,861	0,818	0,818
	2a	0,858	0,834	0,712	0,722	0,774	0,777	0,726	0,740	0,924	0,895	0,828	0,818		
	2b	0,850	0,839	0,631	0,652	0,757	0,756	0,700	0,738	0,925	0,891	0,824	0,776		
33%	2c	0,831	0,853	0,670	0,610	0,759	0,780	0,750	0,744	0,875	0,911	0,767	0,821		
	2d	0,823	0,885	0,668	0,756	0,718	0,829	0,636	0,773	0,925	0,949	0,824	0,895		
	2e	0,839	0,900	0,576	0,763	0,783	0,842	0,818	0,727	0,850	1,000	0,750	1,000		
Ortalama		0,840	0,862	0,651	0,701	0,758	0,797	0,726	0,744	0,900	0,929	0,799	0,862	0,779	0,816
	3a	0,897	0,909	0,667	0,648	0,779	0,777	0,726	0,740	0,954	0,955	0,841	0,818		
	3b	0,852	0,846	0,663	0,650	0,715	0,661	0,733	0,590	0,893	0,933	0,698	0,750		
25%	3c	0,849	0,840	0,646	0,645	0,699	0,667	0,659	0,605	0,917	0,924	0,744	0,743		
	3d	0,889	0,889	0,656	0,599	0,769	0,757	0,682	0,636	0,966	0,983	0,882	0,933		
	3e	0,902	0,878	0,672	0,674	0,800	0,762	0,727	0,727	0,967	0,933	0,889	0,800		
Ortalama		0,878	0,872	0,661	0,643	0,752	0,725	0,705	0,660	0,939	0,946	0,811	0,809	0,791	0,776
	4a	0,913	0,903	0,634	0,608	0,775	0,646	0,753	0,575	0,952	0,963	0,797	0,737		
	4b	0,873	0,893	0,643	0,599	0,655	0,709	0,600	0,639	0,941	0,958	0,720	0,796		
20%	4c	0,891	0,925	0,608	0,564	0,718	0,800	0,636	0,750	0,962	0,969	0,824	0,857		
	4d	0,843	0,863	0,682	0,664	0,579	0,696	0,500	0,727	0,938	0,900	0,688	0,667		
	4e	0,902	0,922	0,611	0,694	0,737	0,800	0,636	0,727	0,975	0,975	0,875	0,889		
Ortalama		0,884	0,901	0,636	0,626	0,693	0,730	0,625	0,684	0,954	0,953	0,781	0,789	0,762	0,781