İktisat Araştırmaları Dergisi • Journal of Research in Economics • Cilt: 3 / Issue • Sayı: 1 Mart • March 2019 • ss/pp. 28-45
ISSN: 2636-8307 • DOI: 10.24954/JORE.2019.27

DİJİTAL REKLAM VERİLERİNDEN YARARLANARAK POTANSİYEL KONUT ALICILARININ RASTGELE ORMAN YÖNTEMİYLE SINIFLANDIRILMASI

CLASSIFICATION OF POTENTIAL RESIDENTIAL BUYERS BY USING RANDOM FOREST METHOD TAKING ADVANTAGE OF DIGITAL ADVERTISING DATA

Haydar EKELİK[*]
Dilek ALTAŞ[**]

Öz

Günümüzde internet ağlarının yaygınlaşması ve internete erişimin bir ihtiyaç haline gelmesi internet sitelerinde ve diğer dijital platformlardaki reklamların kullanılmasını yaygınlaştırmıştır. Dijital reklamcılık olarak adlandırılan bu süreç firmalar, markalar ve diğer kuruluşlar için insanlara ulaşma ve reklam amaçları doğrultusunda hedeflerini gerçekleştirmelerinde gerekli bir reklam aracı olmuştur. En önemli özelliği ölçülebilir olan dijital reklamcılık, firmalara çok geniş veriler (istatistikler) sağlamaktadır. Firmalar bu verileri kullanıp dijital reklamların değerlendirmesini yaparak gelecek reklam planları için ön görüye sahip olurlar. Bu çalışmanın amacı bir inşaat firmasının dijital reklam kampanyasından elde edilen kullanıcı verilerini kullanarak bir sınıflandırma yapmaktır. Kullanıcıların satış ofisine gelip gelmediklerinin kaydının tutulduğu veriler analiz edilerek bir sınıflandırıcı oluşturulmuştur. Bundan sonraki süreçte reklamlarla elde edilen kullanıcı verileri bu sınıflandırıcı kullanılarak sınıflandırılabilir. Böylece kullanıcıların satış ofisine gelip gelmemeleriyle ilgili bir ön bilgi elde edilir. Firma bu ön bilgi sayesinde satış ve pazarlama hedeflerini daha doğru bir şekilde belirleyebilir. Çalışmanın amacı doğrultusunda bağımlı değişken olarak kullanıcıların satış ofisine gelip gelmemesi, bağımsız değişken olarak ise dijital reklamlar sayesinde kullanıcın iletişim bilgilerini hangi gün firma çalışanlarına gönderdiği, kullanıcının cinsiyeti, reklamı hangi sitede görüp siteye geldiği, reklamı hangi reklam alanında (doğal, 300*250 görsel boyutlu vb.) gördüğü, hangi cihazdan (bilgisayar veya telefon) gördüğü, kullanıcının daha önce ilgili firmada kayıtlı olup olmaması ve bu formu hangi amaçla doldurduğu (yatırım, ev sahibi olma vb.) olmak üzere toplamda 7 adet bağımsız değişken kullanılmıştır. Uygulamada R programından yararlanılmış ve verileri analiz etmek için bir topluluk

* İstanbul Üniversitesi İktisat Fakültesi Yöneylem Anabilim Dalı. haydar.ekelik@istanbul.edu.tr, Orcid: 0000-0002-0661-4164

** Marmara Üniversitesi İktisat Fakültesi İstatistik Anabilim Dalı. dilekaltas@marmara.edu.tr, Orcid: 0000-0001-5103-9018

Makale Gönderim Tarihi: 11.02.2019
Yayına Kabul Tarihi: 01.03.2019

öğrenme algoritması olan Rastgele Orman Yöntemi kullanılmıştır. Temelinde karar ağaçları olan bu yöntem diğer sınıflandırma algoritmalarına göre daha iyi sonuçlar vermiştir.

Anahtar Kelimeler: E-ticarette Tahmin Analizi, Kitle Analizi, Rastgele Orman

JEL Sınıflandırması: L81, C45, M3

Abstract

The widespread use of internet networks and the need to access the internet has become widespread in the use of internet sites and other digital platforms. This process, which is called digital advertising, has become an indispensable advertising tool for companies, brands and other organizations to reach their goals and to realize them in accordance with advertising purposes. Digital advertising, the most important feature of which is measurable, gives companies a very large data (statistics). Firms use this data to evaluate the digital advertising and have a look to the future advertising plans. The purpose of this study is to make a classification by using user data from a construction company's digital advertising campaign. A classifier is created by analyzing the data that the users are kept in the sales office and the user data obtained with the subsequent ads can be classified using this classifier. Thus, a preliminary information can be obtained about whether the users come to the sales office. Through this preliminary information, the company will determine its targets in sales and marketing more accurately. Through the purpose of the study, as dependent variable, whether the users came to the sales office, as the independent variable,the user sent which days the contact information to the employees of the company , the gender of the user, the site in which the advertisement is seen , which advertisement area (native, 300*250 visual size etc.), where ads were seen by the user (computer or telephone), whether the user had previously been registered in the relevant company and for which purpose he filled in this form (investment, host, etc.) a total of 7 independent variables were used belirlenmişti. R program was used and the Random Forests Method, a community learning algorithm, was employed to analyze the data. This method, which is based on decision trees, yields better results than other classification algorithms.

Keywords: Predictive Analysis in E-commerce, Mass (Audience) Analysis, Random Forest

JEL Classification: L81, C45, M3

1.Giriş

Son dönemlerde, internet ağlarının yaygınlaşması ve internete erişimin bir ihtiyaç haline gelmesi ile birlikte internet sitelerinde ve diğer dijital platformlardaki reklamların kullanılması yaygınlaşmıştır. Dijital reklamcılık olarak adlandırılan bu süreç, internet teknolojilerinin geldiği son noktada, insanların bilgiye en hızlı ulaştığı, ilk aramayı yaptığı internete bağlı cihazlar üzerinde, markaların ya da ürünlerin tanıtımının yapılmasıdır. Teknolojik gelişmeler ve internete olan erişimin artması sonucunda dijital reklamcılık, reklam sektörü için yeni ve gerekli bir hale gelmiştir(Barfoot, Burtenshaw ve Mahon 2006; Akt: Yürekli,2016). Dijital reklamcılık, günümüzde çevrimiçi platformların (internet site ve mobil uygulamaları) ve içeriklerin en yüksek payını finanse eden baskın bir faaliyet haline gelmiştir(McStay,2010 Yürekli, 2016). Markalar ya da ürünler bu teknolojiyi kullanarak reklam kampanyalarında daha ulaşılabilir ve daha görünür olmaktadır. Günümüzde şirketler, web sitelerini ziyaret ettikten sonra tüketicilerin davranışlarındaki değişikliklere bağlı olarak yapılan araştırmaların sonucunda web sitelerine trafik

sağlamak için çeşitli reklam türleri kullanarak dijital reklamcılık yatırımlarını genişletmektedir. Çoğu şirket, son dönemlerde çevrimdışı reklamlara (televizyon, gazete, radyo vb.) ek olarak farklı türde çevrimiçi reklamlara yatırım yapmaktadır. Şirketler, müşterilerine dijital reklamlar sayesinde kişisel veya belirli bir hedef kitleye özgü reklamları gösterebilirler.

Bu çalışmada bir inşaat firmasının, belirli bir dönemde yapmış olduğu dijital reklam yayınlarından elde edilen veriler (kullanıcı verileri) analiz edilmiştir. İnşaat firmasının reklamlarını dijital platformlarda görmüş olan kullanıcılar bu reklamlar sayesinde firma sitesine gelmiş ve burada konut alımıyla ilgili olarak iletişim bilgilerini firmaya bir form aracılığıyla göndermişlerdir. Bu kullanıcıların (müşteriler) inşaat firmasının satış ofisine gelip gelmedikleri kayıt edilmiştir. Elde edilen bu bilgiler sayesinde kullanıcılar satış ofisine gelip gelmeme durumuna bağlı olarak veri madenciliğinde çokça kullanılan ve karar ağaçları için bir topluluk öğrenme algoritması olan Rastgele Orman (Random Forest) yöntemiyle sınıflandırılmıştır. Böylelikle elde edilen yeni bir kullanıcı bilgisi, geliştirilen model sayesinde bu kullanıcının satış ofisine gelip gelmemesi tahmin edilecektir. Böylelikle firma hangi kullanıcıların satış ofisine geleceğini önceden öngörerek her bir kullanıcıyı arayarak harcadığı zamanı en aza indirecek, hedef kitle analizini gerçekleştirecek ve elindeki kullanıcı verilerini daha etkin şekilde kullanmış olacaktır.

2. Yöntem

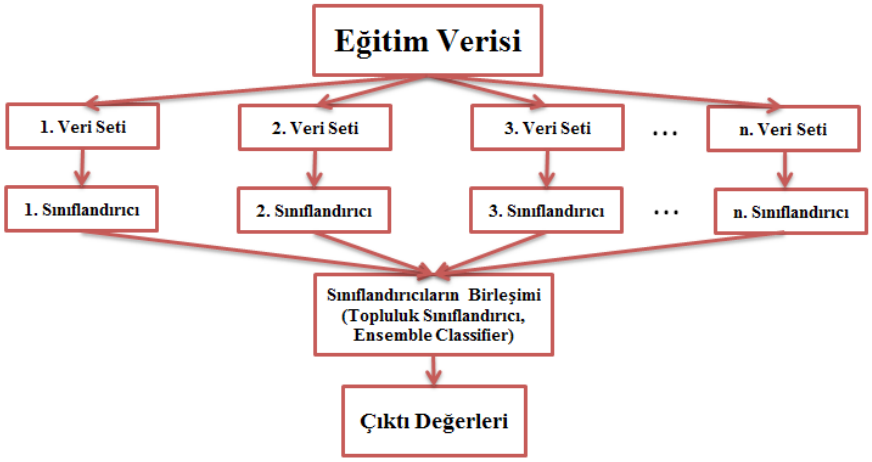
2.1. Karar Ormanları (Topluluk Öğrenme Algoritmaları)

Topluluk öğrenme yöntemlerinin ana fikri, her biri aynı problemi çözen, daha doğru ve güvenilir tahminler veya kararlarla tek bir model kullanarak elde edilenden daha iyi bir model oluşturmaktır. Çoklu modelleri birleştirerek tahmin modelini oluşturma fikri uzun süredir araştırılmaktadır. Topluluk öğrenme yöntemi, önemli bir karar vermeden önce çevresine danışan, farklı görüşler alan bir insanın nihai kararını verme sürecine benzer(Polikar, 2006).

Topluluk yöntemlerinin tahmin performansının iyileştirilmesinde kullanılabileceği bilinmektedir. İstatistik, makine öğrenimi, örüntü tanıma ve veri madenciliği gibi çeşitli disiplinlerden araştırmacılar, topluluk yönteminin kullanımını dikkate almışlardır.

Literatürde, "topluluk yöntemleri" terimi genellikle aynı temel modelin küçük değişikliklerine sahip model koleksiyonları olarak yer almaktadır. Ayrıca, literatürde "çoklu sınıflayıcı sistemler" olarak da bilinmektedir (Rokach ve Maimon, 2014).

Şekil 2.1. Karar Ormanları Bağımssız Sınıflandırıcı Çalışma Prensibi



2.1.1. Torbalama (Bagging)

1996 yılında Breiman tarafından geliştirilen yöntem olan Bootsrap toplaması (bootstrap aggregating) veya Torbalama (bagging), istatistiksel öğrenme yönteminin hata varyansını azaltmak için amaçlanan genel bir yöntemdir. Torbalama tahmini, bir tahmincinin birden çok sürümünü oluşturur ve bunları toplu bir tahmin elde etmek için kullanan bir yöntemdir. Torbalama, sayısal bir sonuç tahmin ederken tahminlerin ortalamasını alır ve bir sınıf tahmin ederken oylamaya göre tahmin yapar(Breiman, 1996). Yöntem, öğrenilmiş sınıflandırıcıların çeşitli çıktılarını tek bir tahmin halinde birleştirerek, geliştirilmiş bir bileşik sınıflandırıcı oluşturarak doğruluğu arttırmayı amaçlamaktadır. Yeni bir örneği sınıflandırmada, her bir sınıflandırıcı bilinmeyen örnek için sınıf tahminini oluşturur. Sonuç olarak, torbalama, orijinal tekli verilerden oluşturulan tek modelden daha iyi performans gösteren birleşik bir model üretmektedir. Breiman 1996 yılındaki makalesinde, özellikle kararsız sınıflandırıcılar için bunun doğru olduğunu belirtmektedir. Çünkü torbalama kararsızlıkları ortadan kaldırabilir. Bu bağlamda, öğrenme setini bozmak, yapılandırılmış sınıflandırıcıda önemli değişikliklere neden olabiliyorsa, kararsız olarak kabul edilir. Bu yöntem karar ağaçları için yararlıdır ve sıklıkla kullanılır(Rokach ve Maimon, 2014).

Torbalama algoritmasını kısaca özetlemek gerekirse; Bir bootstrap örneği ile iadeli olarak eğitim setinden eşit oranda m tane örnek içeren örneklemler üretilir. T boostsrap örneklemleri B_1, B_2, … , B_T üretilir ve her bir bootstrap örneği B_i için bir C_i sınıflandırıcısı oluşturulur. Son sınıflandırıcı olan C^*, C_1,C_2, …,C_T sınıflandırıcılarının en çok tahmin ettiği sınıfı baz alarak elde edilen sınıflandırıcıdır(Bauer ve Kohavi, 1999).

2.1.2. Rastgele Orman (Random Forest)

Rastgele Orman (Random Forest) 2001 yılında Leo Breiman tarafından geliştirilmiştir. Rastgele orman, torbalama yöntemi ve Tim Kam Ho [1] tarafından önerilen rastgele alt uzay yöntemlerinin birleşiminden oluşmaktadır. Rastgele alt uzay yönteminde en uygun dallara bölünmeyi sağlayacak değişken tüm değişkenler arasından rastgele seçilmiş az sayıda değişken tarafından belirlenir(Uzbaş, 2017). Rastgele ormanlar, ağaç tipi sınıflandırıcılar topluluğudur, ormandaki her ağaç bağımsız olarak örneklenen bir rasgele vektörün değerlerine ve aynı dağılıma bağlıdır(Breiman, 1999).Torbalama yönteminin gelişmiş bir şekli olarak kabul edilebilir. Bu yöntemi Breiman makalesinde şöyle açıklamaktadır. *k*.ıncı ağaç için, geçmiş rastgele vektörlerden Q_1, ..., ,Q_{k-1} bağımsız ancak aynı dağılımla sahip Q_k rasgele vektörü oluşturulur; ve bir ağaç, *x*'in giriş vektörü olduğu $h(x,Q_k)$ sınıflandırıcısının sonucu olarak eğitim veri seti ve Q_k kullanılarak oluşturulur(Breiman, 2001). Örneğin, torbalamada, rasgele vektör *Q*, *N* eğitim veri setindeki gözlem sayısı olmak üzere, kutularda rastgele atılan *N* gözlemlerinden elde edilen *N* kutudaki sayımlar olarak oluşturulur. Rastgele bölünme seçiminde *Q*, *1* ile *k* arasında bir dizi bağımsız rastgele tam sayıdan oluşur. *Q*'nun yapısı ve boyutsallığı, ağaç yapımındaki kullanımına bağlıdır(Breiman, 2001).

Rastgele orman yönteminde, torbalama ve rastgele değişken (özellik) seçimi birlikte kullanılır. Her yeni eğitim seti, orjinal eğitim setinden iadeli olarak (bootsrap yöntemiyle) çekilir. Ardından rastgele değişken (özellik) seçimi kullanılarak yeni eğitim setinde bir ağaç yetiştirilir. Yetiştirilen ağaçlarda budama yapılmaz(Breiman, 2001). Yapılan çalışmalar, özellik seçimi ölçütlerinin değil, budama yöntemlerinin seçiminin ağaç tabanlı sınıflandırıcıların performansını etkilediğini göstermektedir(Mingers, 1989). Torbalamanın kullanılmamasının iki nedeni vardır. Birincisi, rastgele değişken seçimi (özellikler) kullanıldığında torbalama kullanımının doğruluğu arttırdığı görülmektedir. İkincisi, torba dışı hataların (Out-of-bag (OOB)) hesaplanmasıdır(Breiman, 2001). Budamanın olmaması rastgele ormanları diğer karar ağacı yöntemlerinden daha avantajlı hale getirmektedir. Torbalama ve rastegele orman yöntemi arasındaki temel fark *m* boyutlu değişkenlerin seçimidir. Örneğin,toplamda *p* değişken varsa ve *m=p* olarak oluşturulursa, o zaman yöntem torbalama'ya eşit olur. $m = \sqrt{p}$ 'yi kullanan rastgele ormanlar hem test hatasında hem de torbalama üzerindeki OOB hatasında bir azalışa yol açar(James vd. 2014).

Rasgele orman karar ağaçları için tanımlanmış olmasına rağmen, bu yaklaşım tüm sınıflandırıcılar için geçerlidir. Rastgele orman yönteminin önemli bir avantajı, çok sayıda girdi değişkenlerini ele almasıdır(Skurichina ve Duin, 2002). Rastgele ormanın bir başka önemli özelliği de hızlı olmasıdır.

1 Ayrıntılı bilgi için bkz. Ho (1998).

3.Verilerin Analizi

Uygulamada bir inşaat firmasının, belirli bir dönemde yapmış olduğu dijital reklam yayınlarından elde edilen veriler (kullanıcı verileri) R programının randomForest kütüphanesi kullanılarak analiz edilmiştir. İnşaat firmasının reklamlarını dijital platformlarda görmüş olan kullanıcılar bu reklamlar sayesinde firma sitesine gelmiş ve burada konut alımıyla ilgili olarak iletişim bilgilerini firmaya bir form aracılığıyla göndermişlerdir. Firma çalışanları kullanıcıların iletişim bilgileri sayesinde onlara ulaşmış ve konutlar hakkında daha ayrıntılı bilgi vermek için uygun bir zamanda satış ofislerine davet etmişlerdir. Tüm bu süreç yani kullanıcının reklamı görmesi, ilgili reklam sayesinde firma sitesine gelmesi, form doldurması ve satış ofisine gelmesi kayıt altına alınmıştır. Uygulama sayesinde bu veri seti analiz edilmiş ve kullanıcılar satış ofisine gelen ve gelmeyen olarak oluşturulan model tarafından tahmin edilmiştir. Amaç bundan sonraki süreçte oluşturulan model yardımıyla yeni kullanıcıların satış ofisine gelip gelmeyeceğinin tahmin edilmesidir. Böylelikle firma hangi kullanıcıların satış ofisine geleceğini önceden öngörerek her bir kullanıcıyı arayarak harcadığı zamanı en aza indirecek, hedef kitle analizini gerçekleştirecek ve elindeki kullanıcı verilerini daha etkin şekilde kullanmış olacaktır.

3.1.Uygulama Kapsamı ve Veri Yapısı

Uygulamada bağımlı değişken olarak kullanıcıların satış ofisine gelip gelmemesi, bağımsız değişken olarak ise kullanıcının iletişim bilgilerini hangi gün firma çalışanlarına gönderdiği, kullanıcının cinsiyeti, reklamı hangi internet sitesinde (mecra) görüp firma sitene geldiği, reklamı hangi reklam alanında (doğal, 300*250 görsel boyutu vb.) gördüğü, hangi cihazdan (bilgisayar veya mobil) gördüğü, kullanıcının daha önce ilgili firmada kayıtlı olup olmaması ve bu formu hangi amaçla doldurduğu(yatırım, ev sahibi olma vb.) olmak üzere toplamda 7 adet bağımsız değişken kullanılmıştır.

Uygulamada kullanılan Rastgele Orman yöntemiyle veri seti analiz edilmiş ve kullanıcılar satış ofisine gelen ve gelmeyen olarak model tarafından tahmin edilmiştir.

Bağımlı değişkene göre eğitim ve test verileri;

Tablo 3.1. Sınıf sayıları

Eğitim Veri Seti;

Satış Ofisi	0	1	Toplam
	397	44	441

Test Veri Seti;

Satış Ofisi	0	1	Toplam
	81	14	95

İlk başta 1.492 kullanıcı verisi elde edilmiş ancak veri temizleme aşamasında geriye 536 kullanıcı verisi kalmıştır. Bu temizleme işlemi kullanıcıların hatalı iletişim bilgisi vermesi, ulaşılamamaları, formu yanlışlıkla doldurmaları ve konut alımıyla ilgilenmedikleri bilgisini vermeleri üzere analizden çıkarılmıştır. Analiz bu 536 kullanıcı verisiyle yapılmıştır. Bu kullanıcı verisinin %80'i eğitim, %20'si test verisi olarak kullanılmıştır. Bağımlı değişken olan satış ofisine gelme durumunu gösteren değişken, gelenler 1, gelmeyenler 0 sembolleriyle gösterilmiştir.

Uygulamada kullanılan bağımsız değişkenlerin hepsi kategoriktir.

Gün değişkeni; 1'den 7'ye kadar olan rakamlarla gösterilmiştir. 1 pazartesi gününe karşılık gelmekte olup sırayla devam ederek 7 Pazar gününe karşılık gelmektedir. Kullanılan R programında gun diye kodlanmıştır.

Cinsiyet değişkeni; 0 ve 1 rakamlarıyla gösterilmiştir. 0 erkeklere 1 kadınlara karşılık gelmektedir. Kullanılan R programında cins diye kodlanmıştır.

Cihaz değişkeni; 0 ve 1 rakamlarıyla gösterilmiştir. 0 mobil(cep telefonu) 1 bilgisayara karşılık gelmektedir. Kullanılan R programında cih diye kodlanmıştır.

Mecra değişkeni;1'den 23'e kadar olan rakamlarla gösterilmiştir. 1. mecra, 2. mecra olarak 23'e kadar numaralandırılmıştır. Kullanılan R programında mec diye kodlanmıştır.

Kreatif değişkeni; 1'den 21'e kadar olan rakamlarla gösterilmiştir. 1. kreatif, 2. kreatif olarak 21'e kadar numaralandırılmıştır. Kullanılan R programında kre diye kodlanmıştır.

Kayıt değişken; 0 ve 1 rakamlarıyla gösterilmiştir. 0 kayıtlı olmayan kullanıcıya, 1 kayıtlı kullanıcıya karşılık gelmektedir. Kullanılan R programında kay diye kodlanmıştır.

Görüş değişken; 1'den 6'ya kadar olan rakamlarla gösterilmiştir. Kullanılan R programında grs diye kodlanmıştır.

1; 1+1 konutlarla ile ilgilenen kullanıcı

2; 2+1 konutlarla ilgilenen kullanıcı

3; 3+1 konutlarla ilgilenen kullanıcı

4; detaylı bilgi almak isteyen kullanıcı

5; yatırım yapmak isteyen kullanıcı

6; diğerler amaçlar için, olarak ifade edilmiştir.

Bağımsız değişkenlerin 0 ve 1 arasında olması ve değişken normlarının eşit olması için gün, mecra, kreatif ve görüş değişkenlerine aşağıdaki denklemde gösterildiği gibi normalizasyon dönüşümü yapılmıştır(Han, Kamber ve Pei, 2012).

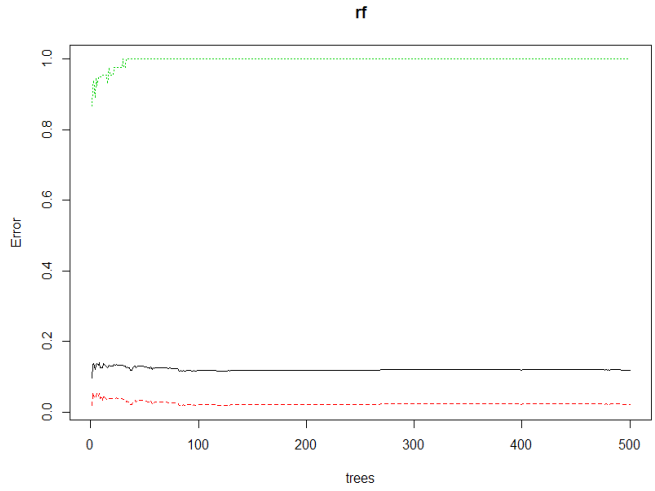
$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (3.1)$$

3.2.Elde Edilen Bulgular

Uygulama analiz sonuçları, R (R Core Team,2017) programında randomForest(Liaw ve Wiener,2002) paketi kullanılarak elde edilmiştir.

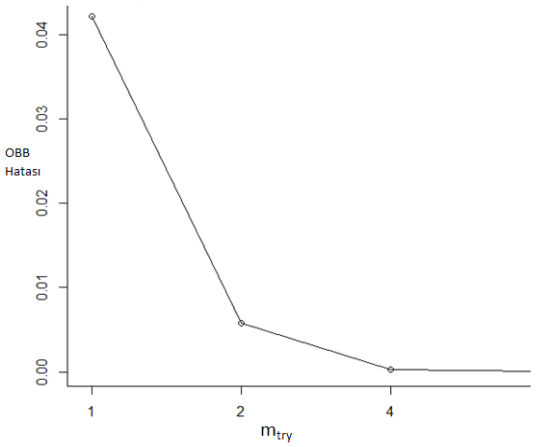
Öncelikle ağaç sayısı ve ağaç oluşumda her bölünmede dikkate alınan değişkenlerin sayısı keyfi olarak belirlenir ve yapılan analiz sonucunda bu değerler için optimum değerler bulunarak bu parametreler yeniden belirlenir. Keyfi olarak ormanda oluşan ağaç sayısının 500 seçildiği ve ağaç oluşumda her bölünmede dikkate alınan değişkenlerin yine keyfi olarak 2 olarak seçildiği yöntemin çıktıları şekil 3.1 ve şekil 3.2'de gösterilmiştir.

Şekil 3.1. Ağaç sayısı grafiği



Grafik OOB hatasının ağaç sayısına bağlı olarak grafiğini çizmektedir. Ağaç sayısının artıkça bu hatanın sabitleştiğini görülmektedir. Yeni oluşturulacak parametre değerinde ağaç sayısı 200 olarak alınabilir ya da 500 olarak kalabilir. Ağaç sayısının fazla olması OOB'de bir azalışa yol açmasa da ormanın daha kararlı olmasını sağlar (James ve diğerleri, 2014).

Şekil 3.2. Ağaç oluşumda her bölünmede dikkate alınan değişken sayısı



OBB hatasını en aza indirecek her bölünmede dikkate alınan değişken sayısının grafiğine bakarak bu parametre değerinin 4 olması gerektiği söylenebilir.

Kullanıcı tabanlı parametrelerin optimal değerinin ne olacağı belirlendikten sonra analiz tekrardan yapılır ve aşağıdaki sonuçlar elde edilir.

Tablo 3.2. Eğitim verisi için sınıflandırma tablosu

	Gerçek	
Tahmin	**0**	**1**
0	**395**	**17**
1	**2**	**27**

Tablo 3.3. Eğitim verisi için sınıflandırma tablosuna ait değerlendirme ölçütleri

Doğruluk=	**95,69%**
Yanlış Sınıflandırma Oranı=	**4,31%**
Kesinlik=	**93,10%**
Doğru Pozitif Oran=	**61,36%**
Yanlış Pozitif Oran=	**4,55%**
Doğru Negatif Oran=	**99,50%**
Yanlış Negatif Oran=	**38,64%**
Kappa=	**71,73%**

Kappa istatistiği; Cohen'in kappa katsayısı olarak bilinir ve iki değerleyici arasındaki karşılaştırmalı uyuşmanın güvenirliğini ölçen bir istatistik yöntemidir. Eğer Pr(*a*) iki değerleyici için örtüşen uyuşmaların toplama olan oranı ve Pr(*e*) ise bu uyuşmanın şans eseri ortaya çıkma olasılığı ise, Cohen'in kappa katsayısı bulunması için kullanılacak formül şu olur: Aşağıdaki formül yardımıyla hesaplanır. Kappa değerinin 1'e yakın olması iki değerleyicinin o derece iyi bir şekilde uyuşmakta olduğunu göstermektedir (Cohen,1960).

$$\kappa = \frac{\Pr(\alpha) - \Pr(e)}{1 - \Pr(e)} \qquad (3.2)$$

Tablo 3.3'de görüldüğü gibi eğitim verisinde doğruluk %95 gibi çok yüksek bir değer çıkmıştır. Doğru pozitif oranın %61,36 olduğu görülmektedir. Bu değerler karar ormanın eğitim verisini iyi anlamda öğrendiğini göstermektedir. Kappa istatistiğinin %70 olması da modelin yüksek oranda gerçek değerlerle uyuştuğunun bir göstergesidir.

Tablo 3.4. Test verisi için sınıflandırma tablosu

	Gerçek	
Tahmin	**0**	**1**
0	**77**	**13**
1	**4**	**1**

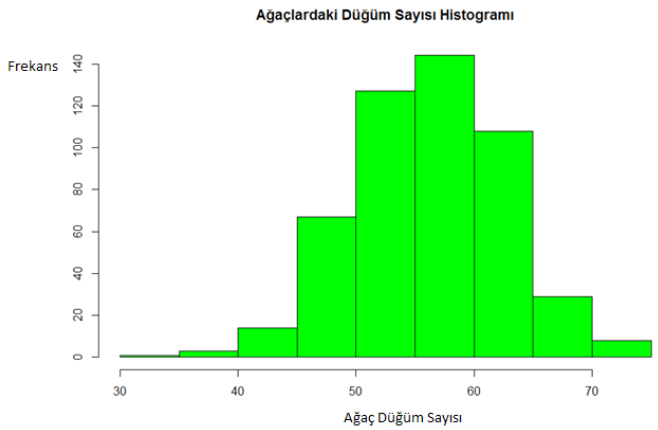
Tablo 3.5. Eğitim verisi için sınıflandırma tablosuna ait değerlendirme ölçütleri

Doğruluk=	**82,11%**
Yanlış Sınıflandırma Oranı=	**17,89%**
Kesinlik=	**20,00%**
Doğru Pozitif Oran=	**7,14%**
Yanlış Pozitif Oran=	**28,57%**
Doğru Negatif Oran=	**95,06%**
Yanlış Negatif Oran=	**92,86%**
Kappa=	**0,05%**

Tablo 3.5'de görüldüğü gibi test verisinin doğruluğunun yüksek %82 gibi yüksek bir değer olmasına rağmen doğru pozitif oranının %7 olması veri yapısından kaynaklanmaktadır. Bu durumu Brieman ve arkadaşları bir makalesinde şöyle açıklamaktadır. Random forest sınıflandırıcısı, çoğunluk sınıfını azınlık sınıfına göre daha doğru sınıflandırma eğilimi gösterir. Bu nedenle

azınlık sınıfını iyi bir şekilde sınıflandıramaz. Bu durumda ağırlıklandırılmış Rf (Random Forest) algoritması kullanılır. Ancak yapılan çalışmalar dengesiz veriyi öğrenmek için kullanılan ağırlıklı Rf ile dengeli Rf (uygulamada kullanılan) arasında net bir kazananın olmadığını ortaya çıkarmıştır (Chen vd., 2003).

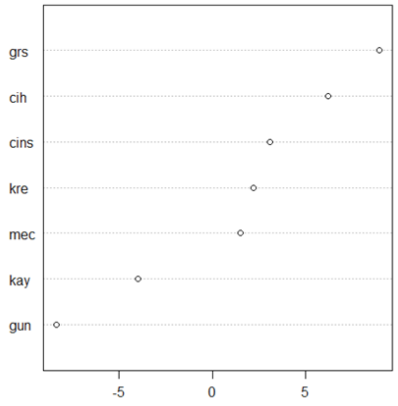
Şekil 3.3. Ağaçlardaki düğüm sayısı histogramı



Toplamda 501 tane ağaçtan oluşan ormanda düğüm sayılarının histogramı yukarıda verilmiştir. Ağaçlardaki düğüm sayısı yukarıdaki grafikten görüldüğü gibi ortalama olarak 50-60 arasında değişmektedir.

Ortalama Düşüş Doğruluğu

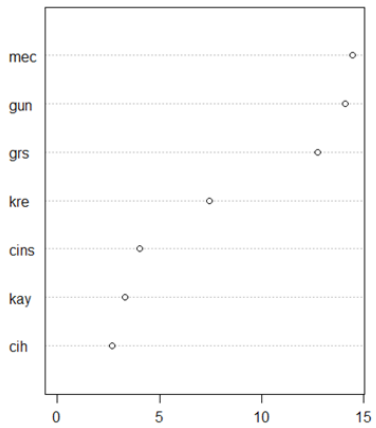
Şekil 3.4. Değişkenlerin doğruluğa olan katkıları



Şekil herhangi bir değişken olmadığında model performansının ne kadar kötü olduğunu test eder. Ağaçları yaparken herhangi bir değişkenini kaldırırsak, doğrulukta ortalama azalma ne olur sorusunun cevabını veren grafiktir. Örneğin grs (görüş) değişkeninin değeri çok yüksek olduğundan doğruluğun hesaplanmasında yüksek bir öneme sahip olduğu söylenir.

Gini Değeri Ortalama Düşüşü

Şekil 3.5. Düğümlerin değişkenlere göre saflık değerleri



Bu grafik, herhangi bir değişken olmadan ağacın sonunda düğümlerin ne kadar saflıkta olduğunu ölçer. Yani değişkenlerden biri modelden çıktığında Gini katsayında ne kadar bir azalma olacağının cevabını verir. Örneğin mec (mecra) değişkeni değeri yüksek olduğu için diğer değişkenlere göre Gini indeksinde yüksek katkısı olduğu söylenir.

Tablo 3.6. Değişkenlerin doğruluk ve gini katkıları için sayısal değerler

Değişkenler	Ortalama Doğruluk	Gini
gun (gün)	-8,36	14,09
cins (cinsiyet)	3,11	4,03
cih (cihaz)	6,22	2,67
mec (mecra)	1,49	14,46
kre (kreatif)	2,2	7,43
grs (görüşme)	8,97	12,75
kay (kayıt)	-4	3,31

Bu tablo yukarıdaki verilen grafiklerde yer alan değişkenlerin sayısal değerlerini göstermektedir.

Tablo 3.7. Değişken kullanım sayıları

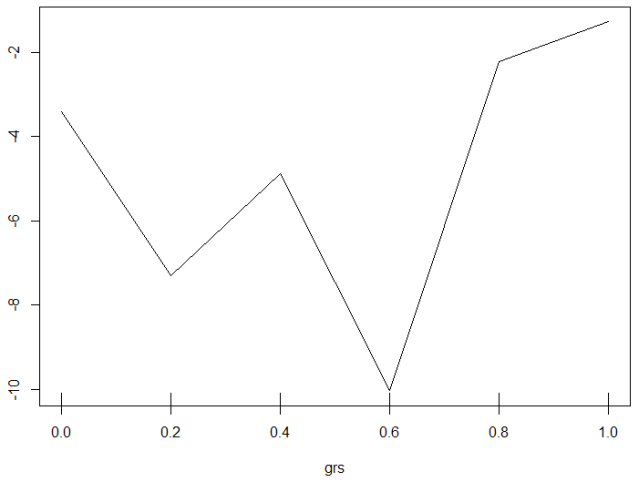
Değişkenler	Kullanım Sayıları
gun (gün)	7.760
cins (cinsiyet)	2.215
cih (cihaz)	873
mec (mecra)	7.119
kre (kreatif)	4.120
grs (görüşme)	3.534
kay (kayıt)	2.275

Bu tablo, değişkenlerin ağaç oluşumdaki kullanım sayılarını göstermektedir. Örneğin en fazla kullanılan değişken gun(gün) değişkenidir. Bu değişkenin ortalama doğruluğa katkısı – 8,36 iken gini indeksine katsısı ise 14,09 olarak hesaplanmıştır.

Kısmi Bağımlılık grafikleri;

Bu grafikler bir değişkenin, sınıf olasılığı üzerindeki marjinal etkisini vermektedir.

Şekil 3.6. Görüş değişkenin kısmi bağımlılığı



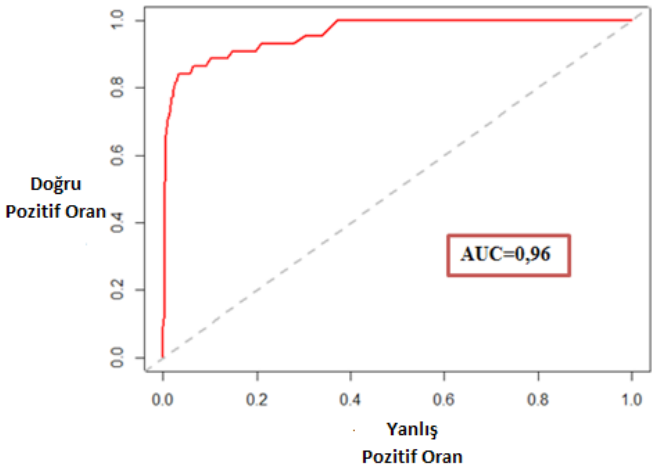
Şekil 3.6 yorumlandığında grs (görüş) değişkeni 0,6 değerinden büyük olduğunda 1 sınıfını 0,6 değerinden küçük olmasına göre daha kuvvetli tahmin etmektedir.

Diğer değişkenlerin kısmi bağımlılıkları da o değişkenlere göre çizilen grafiklere bakılarak yapılabilir. Bu grafikler ekler kısmında verilmiştir.

ROC Eğrileri

Kullanılan random forest algoritması gözlemleri aldığı oylara göre sınıf atamaları yapmaktadır. Bu oylar her bir gözlem için hesaplanır ve OOB örnek ağacı sınıflandırması için de hesaplanır. Bu oylar kısaca bir olasılığı temsil eder ve bu nedenle ROC grafiği çizilip AUC değeri hesaplanabilir. [2]

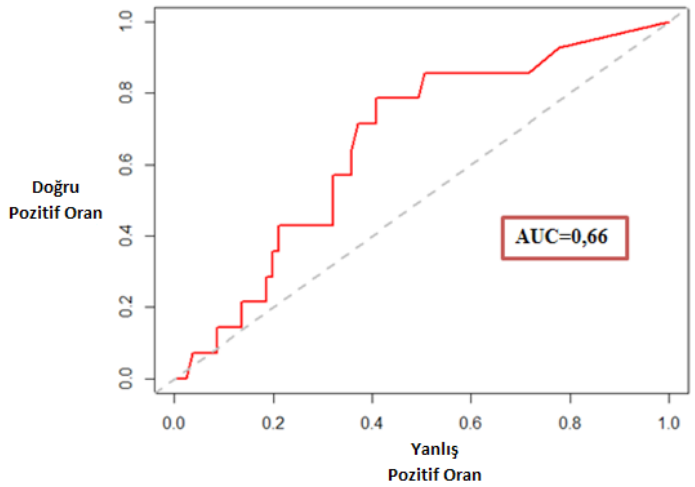
Şekil 3.7. Eğitim verisi için ROC eğrisi



R programındaki ROCR paketi kullanılarak çizilen ROC eğrisi Şekil 3.7'de gösterilmektedir. Yine aynı paket kullanılarak hesaplanan AUC (area under curve) değeri **0,96** olarak hesaplanmıştır. Bu değer sınıflandırma performansının çok iyi olduğunu göstermektedir.

2 www.stats.stackexchange.com/questions/188616/how-can-we-calculate-roc-auc-for-classification-algorithm-such-as-random-forest,2017 Erişim tarihi:12.11.2018

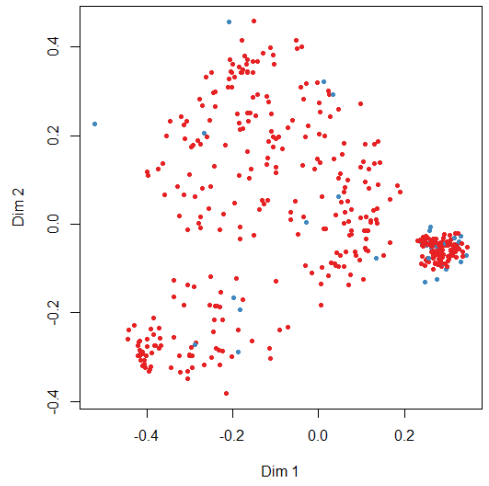
Şekil 3.8. Test verisi için ROC eğrisi



Test verisi için hesaplanan ROC eğrisi Şekil 3.8'de gösterilmiştir. Bu eğri altında kalan AUC değeri de 0,66 olarak bulunmuştur. Bu değer 0,5'nin üzerinde olduğu için sınıflandırma performansısın iyi olduğu söylenebilir.

Çok Boyutlu Ölçekleme Grafiği

Şekil 3.9. Eğitim verisindeki geliş değişkenin çok boyutlu ölçekleme grafiği



Veride 0 sınıfına ait olan gözlemler çoğunlukta olduğu için kırmızı renkli olanlar 0 sınıfını, mavi renkli olanlar ise 1 sınıfını göstermektedir.

4.Sonuç

Çalışmada bir inşaat firmasının dijital reklam yayınları sonucunda elde edilen kullanıcı verileri analiz edilmiştir. Veri analizinde Rastgele Orman Yöntemi kullanılarak sınıflandırma modeli oluşturulmuştur. Kullanıcıların satış ofisine gelip gelmemesinin bağımlı değişken olduğu sınıflandırma modelinde, kullanıcının iletişim bilgilerini hangi gün firma çalışanlarına gönderdiği, kullanıcının cinsiyeti, reklamı hangi internet sitesinde görüp firma sitesine geldiği, reklamı hangi reklam alanında gördüğü, hangi cihazdan (bilgisayar veya telefon) gördüğü, kullanıcının daha önce ilgili firmada kayıtlı olup olmaması ve bu formu hangi amaçla doldurduğu (yatırım, ev sahibi olma vb.) olmak üzere toplamda 7 adet bağımsız değişken kullanılmıştır.

Bağımsız değişkenlerin normlarını eşitlemek için tüm bağımsız değişkenleri 0-1 aralığına dönüştüren normalizasyon dönüşümü yapılmıştır. Oluşturulan sınıflandırma modelinde eğitim verisinde sınıflandırma doğruluğu %95 seviyesinde yüksek bir oranda olurken test verisinde bu oran %82 seviyelerinde gerçekleşmiştir. Bu tarz çalışmalarda test verisi dikkate alınarak değerlendirme yapılmaktadır. Test verisi içinde doğruluk iyi bir seviyede gerçekleşmiştir.

Değişken bazında en fazla görüş değişkeni olmak üzere, cihaz değişkeninin de doğruluğu artırıcı bir etkisi olduğu ortaya çıkmıştır. Yani kullanıcıların (müşterilerin) satış ofisine gelmelerindeki en önemli değişken (tahmin doğruluğunu arttıran değişken) görüş değişkeni ve cihaz değişkeni olduğu görülmektedir. Bu değişkenler tahmin doğruluğuna en fazla katkıyı veren değişkenlerdir. Bu sonuç bundan sonraki form doldurma sayfalarında kullanıcıların konut alma amaçlarını yansıtacak bir seçeneğin olması gerektiğinin ön bilgisini vermektedir. Mecra ve gün değişkenlerinin gini indeksine en fazla katkıyı sağlayan değişkenler olduğu görülmüştür. Mecra ve gün değişkenleri ormandaki karar ağaçlarında en fazla kullanılan değişkenler olmuşlardır. Ayrıca oluşturulan karar ağaçları ortalama 50-60 düğümden oluşmaktadır.

Eğitim verisinde, 0 etiketine sahip olan sınıfın doğru sınıflandırılmasını gösteren doğru negatif oran (duyarlılık) değeri %99 seviyelerinde, 1 etiketine sahip olan sınıfın doğru sınıflandırılmasını gösteren doğru pozitif oran (özgüllük) değeri %61 gibi iyi bir seviyede gerçekleşmiştir. Test verisinde doğru negatif oran (duyarlılık) değeri %95 seviyesinde ve doğru pozitif oran (özgüllük) %7 seviyelerinde olmuştur. Test verisinde özgüllük değerini küçük ve duyarlılık değerinin yüksek çıkma nedeni azınlık sınıf varlığından kaynaklanmaktadır. Yani 0 olarak etiketlenen sınıftaki gözlem sayısı, 1 olarak etiketlenen sınıftaki gözlem sayısından çok fazladır. Bu tip verilerde çoğunluk sınıfı iyi bir seviyede tahmin edilirken azınlık sınıfın tahmin doğruluğu düşük seviyelerde gerçekleşmektedir

Kullanılan Random Forest algoritmasında gözlemlerin aldıkları oylara göre hangi sınıfta oldukları belirlenir. Bu oylar bir olasılığı temsil etmektedir. Böylece ROC grafiği çizilip AUC değeri

hesaplanabilir. Çizilen ROC eğrilerinde eğitim verisi için eğri sol üst köşeye yakın olup eğri altında kalan alan ise 0.96'dir. Eğri altında kalan alanın 1'e yakın olması yapılan sınıflandırmanın iyi seviyelerde olduğunu göstermektedir. Test verisi için çizilen ROC eğrisi sol üst köşeye fazla yakın olmamasına rağmen eğri altında kalan alan 0.66 seviyelerinde olmuştur. Buradan da eğri altında kalan alan 0.5'in üzerinde olduğu için yapılan sınıflandırmanın rastlantısal olmadığı sonucuna varılır. Hem eğitim verisi hem de test verisi değerleine bakarak yapılan sınıflandırmanın iyi bir seviyede olduğu söylenebilir.

Oluşturulan sınıflandırma modeli ile veri tabanına gelen yeni bir kullanıcı bu modele göre belirli hata oranlarıyla sınıflandırılabilir. Böylece yeni kullanıcılar hakkında önceden bir ön bilgiye sahip olunur. Bu ön bilgi sayesinde firma her bir kullanıcıyı arayarak harcadığı zaman kaybını minimize edebilir ve müşteri portföyünü bu bilgiler ışığında değerlendirme imkânına kavuşur. Değişken bazında en önemli değişkenlerin mecra, gün ve görüş değişkenleri olduğu görülmüştür. Bu önemli değişkenlerin kısmi grafiklerine bakılarak yeni hedef kitleler belirlenebilir ve bu hedef kitlelere göre dijital reklam stratejileri oluşturulabilir. Görüş değişkenin önemli bir değişken olarak bulunması sonucunda bundan sonraki form doldurma sayfalarında kullanıcıların konut alma amaçlarını yansıtacak bir seçeneğin olmasını gerektiğinin ön bilgisini de vermektedir**.** Bir başka kullanım amacı ise, şu anki mevcut durumda satış ofisine gelmeyen ancak model tarafından satış ofisine geldi olarak tahmin edilen kullanıcılar da belirlenip bu kullanıcılara özel hedeflemeler ve satış ofisine gelmelerini sağlayacak farklı pazarlama stratejileri de geliştirilebilir.

Referanslar

Bauer, E. ve Kohavi, R. (1999). An Empirical Comparison of Voting Classification Algorithms : Bagging , Boosting , and Variants. Machine Learning, 36: 105–139.

Breiman, L. (1996). Bagging Predictors. Machine Learning, 24: 123–140.

Breiman, L. (1999). Random Forests – Random Features, Tecnical Report 567, Statistic Department, University of California, Berkeley, 1–29. (https://www.stat.berkeley.edu/~breiman/random-forests.pdf, erişim tarihi:08.10.2018).

Breiman, L. (2001). Random forests. Machine Learning, 45(1): 5–32. https://doi.org/10.1023/A:101.093.3404324

Chen, C., Liaw, A. ve Breiman, L. (1999). Using Random Forest to Learn Imbalanced Data, 1–12.

Cohen, J. (1960). A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20(1): 37-46

Han, J., Kamber, M. ve Pei, J. (2012). Data Mining Concepts and Techniques Third Edition. Simon Fraser University: Morgan Kaufman Publıshers.

Ho, T. K. (1998). The Random Subspace Method for Constructing Decision Forests, IEEE Trans. Pattern Anal. Mach. Intell. 20: 832-844.

James, G., Witen, D., Hastie, T. ve Tibshirani, R. (2014). An Introduction to Statistical Learning with Applications in R,. Performance Evaluation, 64. https://doi.org/10.1016/j.peva.2007.06.006

Liaw,A. ve Wiener, M. (2002). Classification and Regression by randomForest. R News 2(3),18-22.

Mingers, J. (1989). An empirical comparison of selection measures for decision tree induction. Machine Learning, 3: 319–342. https://doi.org/10.1007/BF00116837

Polikar, R. (2006). Ensemble based systems in decision making. IEEE Circuits and Systems Magazine, 6(3): 21–44. https://doi.org/10.1109/MCAS.2006.168.8199.

R Core Team (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/.

Rokach, L. ve Maimon, O. (2014). Data Mining With Decision Tree Theory and Applications, 2.Edition, World Scientific Publishing.

Skurichina, M. ve Duin, R. P. W. (2002). Bagging, boosting and the random subspace method for linear classifiers. Pattern Analysis and Applications, 5(2), 121–135. https://doi.org/10.1007/s100.440.200011.

Uzbaş, B.(2017), Sayısal Dental Modellerden Otomatik Cinsiyet Tespiti, Konya (Doktora Tezi).

Yürekli,K..(2016). Dijital Reklamcılıkta Reklam Ajansı – Reklam Veren İlişkisinin Analizi, İstanbul, (Yüksek Lisans Tezi)