Aktüerya Derneği

İstatistikçiler Dergisi:
İstatistik&Aktüerya 7 (2014) 1-13

İstatistikçiler Dergisi:
İstatistik&Aktüerya

Doğrudan Pazarlama Stratejilerinin Belirlenmesinde Veri Madenciliği Yöntemlerinin Kullanımı

Muhsin Özgür Dolgun

Anadolu Araştırma
Bestekar Sokak 42/1 Kavaklıdere
06680 Çankaya, Ankara, Türkiye
ozgurdolgun@anadoluarastirma.com

Derya Ersel

Hacettepe Üniversitesi
İstatistik Bölümü
06800 Çankaya, Ankara, Türkiye
dtektas@hacettepe.edu.tr

Özet

Doğrudan pazarlama, ürünlerin olası müşterilerinin belirlenmesi ve belirlenen müşteri kitlesine bu ürünlerin tanıtımının yapılması sürecidir. Son zamanlarda, genel kitlelere yönelik pazarlama kampanyalarının çok başarılı olmaması nedeniyle firmalar kitle pazarlama yöntemleri yerine hedef kitleye yönelik doğrudan pazarlama yöntemlerine daha çok önem vermektedir. Özellikle baskı ve rekabetin yoğun bir şekilde yaşandığı bankacılık sektöründe doğrudan pazarlama yöntemlerinin başarı oranının daha yüksek olduğu görülmüştür. Veri madenciliği yöntemleri, doğrudan pazarlama kampanyalarına etki eden faktörleri belirleyerek kampanyaların başarısının arttırılmasında kullanılır. Böylece, mevcut kaynakların daha iyi yönlendirilmesini ve potansiyel müşterilerin makul ve doğru bir kümesinin oluşturulmasını sağlar. Bu çalışmada, karar ağaçları, lojistik regresyon, Bayesci ağlar ve destek vektör makineleri gibi veri madenciliği yöntemleri kullanılarak bankacılık sektöründe doğrudan pazarlama kampanyalarının nasıl yönlendirilebileceği üzerinde durulmuştur. Ayrıca, bu tür verinin çözümlenmesinde sıkça karşılaşılan bir durum olan "dengesizlik" problemi incelenmiştir. Çalışmanın sonucunda, genel başarı ölçütüne göre sırasıyla, SVM linear, lojistik regresyon ve SVM RBF, F ölçütüne göre sırasıyla, Lojistik regresyon, SVM RBF ve CHAID ve matthews korelasyon katsayısına göre sırasıyla, SVM linear, lojistik regresyon ve CHAID yöntemleri en başarılı yöntemler olarak tespit edilmiştir.

Anahtar sözcükler: Doğrudan pazarlama; Veri madenciliği; Sınıflandırma; Dengesizlik.

Abstract

Use of Data Mining Methods in Determining Direct Marketing Strategies

Direct marketing is the process of identifying possible customers of products and promoting these products to this specified customer mass. Recently, due to the fact that mass marketing campaigns targeting general public are not successful, firms give more importance to direct marketing campaigns targeting a specific set of customers. Direct marketing methods are more successful escpecially in banking sector where there is more pressure and competition according to other sectors. Data mining methods are used to increase the success of direct marketing campaigns by identifying the factors that effect these campaigns. Thus, these methods provide to direct available resorces and to create a reasonable and true set of potential customers. In this study, we focus on how direct marketing campaigns can be directed in banking sector by using data mining methods such as decision trees, logistic regression, Bayesian networks and support vector machines. Also, we examine class imbalance problem which frequently encountered in the analysis of this kind of data. As a result, SVM linear, logistic regression and SVM RBF methods were the most successful methods according to the overall accuracy metric. Moreover, according to the F measure, logistic regression, SVM RBF and CHAID, and according to the matthews correlation coefficient, SVM linear, logistic regression and CHAID methods have been identified as the most successful methods, respectively.

Keywords: *Direct marketing; Data mining; Classification; Imbalance.*

1. Giriş

Ürün ya da hizmet satışının yapıldığı sektörlerde, tüm firmalar bu ürün ve hizmetlerin tanıtımını yapma ihtiyacı duyarlar. Literatürde, tanıtım ve reklam için "kitlesel pazarlama" ve "doğrudan pazarlama" olmak üzere iki yaklaşım söz konusudur. Gazete, radyo, televizyon gibi kitle iletişim araçlarından yararlanan kitlesel pazarlamada, ilgili ürün ve hizmetin tanıtımı ayrım yapılmaksızın tüm halka yapılır. Bu yaklaşım, tanıtımı yapılan ürüne/hizmete tüm halk tarafından yoğun bir talep olduğunda etkilidir [11]. Bununla birlikte, piyasa şartlarının ağır ve rekabetin çok yoğun olduğu günümüz dünyasında kitlesel pazarlama etkisini yitirmiştir. Tanıtımına yüksek bütçe ayrılan ürün ve hizmetin, bu tanıtımın ulaştığı kişiler tarafından satın alınma oranı % 1'lere kadar düşmüştür [11, 12, 13]. Bu nedenle, özellikle, bankacılık, finans, sigortacılık ve telekomünikasyon gibi sektörlerde kitlesel pazarlama yerine doğrudan pazarlama yöntemleri daha çok kullanılmaya başlamıştır [13].

Doğrudan pazarlamada, hiçbir ayrım yapılmaksızın tüm halka ürün/hizmet tanıtımı yapmak yerine, kişilerin özellikleri ve ihtiyaçları belirlenerek, o ürüne/hizmete ihtiyacı olabilecek ya da satın alma potansiyelinin daha yüksek olduğu belli bir hedef kitleye tanıtım yapılır. Böylece tanıtım kampanyasına geri dönüş oranı arttırılmaya çalışılır [11]. Doğrudan pazarlamada temel hedef, bireysel müşterilerle birebir, iki yönlü, düşük maliyetli iletişim kurmaktır. Bunun için de mevcut müşteri profilini ortaya çıkarmak ve gelecekteki müşteri tercihlerini tahmin etmek gerekir. Günümüzün düzensiz piyasa koşullarında müşteri tercihleri dinamik olarak değişmekte ve doğrudan ortaya çıkartılması zorlaşmaktadır [13].

Günümüzde, müşteri bilgileri büyük veri tabanlarında saklanmaktadır. Böylece, büyük veri kümelerinin çözümlenmesinde kullanılan veri madenciliği yöntemleri doğrudan pazarlama stratejilerinin belirlenmesinde kullanılabilir. Doğrudan pazarlamada müşteri davranış örüntüleri, veri madenciliği yöntemleri ile analiz edilerek ürün tanıtımının yapılacağı potansiyel müşteri profili belirlenebilir [11, 13]. Veri madenciliği, büyük veri tabanlarından bilgi ortaya çıkarmak için makine öğrenmesi, örüntü tanıma, istatistik, veri tabanı, görüntüleme gibi bilim dallarından birçok tekniği bir arada kullanan disiplinler arası bir bilim dalıdır [14]. Bu çalışmada, bankacılık sektöründe doğrudan pazarlama stratejilerinin belirlenmesinde veri madenciliğinden yararlanılmıştır ve daha önce Moro ve arkadaşlarının 2011 yılında yaptıkları bir çalışmada ele aldıkları bir doğrudan pazarlama kampanyasına ait veri kümesi üzerinde analizler gerçekleştirilmiştir. Veri kümesi, Portekiz'deki bir bankanın gerçekleştirdiği doğrudan pazarlama kampanyasında Mayıs 2008 – Kasım 2010 tarihleri arasında telefon vasıtasıyla topladığı bilgiyi içermektedir. Burada amaç, Çizelge 1'de verilen 16 bağımsız değişkenden yararlanılarak müşterilerin vadeli mevduat hesabı açtırıp açtırmayacaklarını tahmin etmektir [12]. Çalışmada, karar ağaçları, lojistik regresyon, Bayesci ağlar ve destek vektör makineleri gibi sınıflama yöntemleri kullanılmış ve elde edilen sonuçlar karşılaştırılmıştır. Bankacılık kampanyalarında, kullanılan yöntem doğrudan pazarlama olsa dahi geri dönüş oranı yine çok yüksek olmadığı için bağımlı değişkende genellikle dengesiz bir dağılım söz konusudur. Bu durum sınıflama yöntemlerinin sonucunu etkileyeceği için verinin dengeli bir yapıya getirilmesi gerekmektedir. Bu çalışmada ayrıca, bu sorun çözümlenerek sağlıklı sonuçlar elde edilmeye çalışılmıştır.

Çalışmanın ikinci bölümünde, veri madenciliği süreci üzerinde genel olarak durulmuş ve kullanılan sınıflama yöntemlerinden kısaca bahsedilmiştir. Üçüncü bölümde, doğrudan pazarlamada müşteri profilinin ortaya çıkartılması için veri madenciliği sürecinden nasıl faydalanılacağı ve karşılaşılan problemler üzerinde durulmuştur. Dördüncü bölümde ise Çizelge 1'deki veri kümesi üzerinde analizler gerçekleştirilmiş ve sonuçlar karşılaştırılmıştır. Çalışmanın beşinci bölümünde ise elde edilen genel sonuçlar yorumlanmış ve karşılaştırma sonuçları tartışılmıştır.

2. Veri Madenciliği

Veri madenciliği, büyük veri kümelerinin, istatistik, matematik ya da örüntü tanımlama teknikleri yardımıyla incelenerek bu veriden yeni ilişkilerin, örüntülerin ve trendlerin keşfedilmesi sürecidir [10]. Veri madenciliğinde, üzerinde çalışılan veri kümesi büyüktür. Küçük veri kümesi üzerinde çalışılırsa, veri

madenciliğinin klasik istatistiksel veri çözümlemesinden bir farkı kalmaz. Veri kümesi büyük olduğunda ise, veriye nasıl ulaşılacağı, verinin nasıl saklanacağı, verinin nasıl çözümleneceği gibi yeni problemler ortaya çıkar [7]. Bu bölümde ilk olarak, veri madenciliği problemlerinin çözümlenmesinde kullanılan bir metodoloji olan Çapraz Endüstri Standart Süreci (The Cross-Industry Standard Process for Data Mining / CRISP-DM) üzerinde durulacaktır. Daha sonra, veri madenciliğinde kullanılan sınıflama modelleri kısaca açıklanacaktır.

2.1. CRISP-DM Süreci

Veri madenciliğinin birçok disiplini barındıran yapısı ve farklı uygulama alanlarındaki görevlerle prosedürlerin çeşitliliği, veri büyüklüğünden dolayı farklı ve kirli veri kaynakları ile çalışmadaki zorluklardan dolayı standart bir metodolojiye ihtiyaç duyulmaktadır. Proje öncesinde aşağıdaki soruların cevaplanması projenin başarısı ve planlanması için yararlı olacaktır.

- Nasıl bir iş problemi çözülmeye çalışılıyor?
- Hangi tür veri kaynakları mevcut ve bu iş problemi için ne çeşit veriye ihtiyaç duyulacak?
- Veri analiz edilmeden önce ne tür ön işleme ve veri temizleme işlemleri gerçekleştirilecek?
- Hangi veri madenciliği yöntemleri kullanılacak?
- Sonuçlar nasıl değerlendirilecek?

Bu tür sorular ve kullanılması yararlı olan bir metodoloji ile veri madenciliği süreci içerisindeki karmaşık yollarda kolayca ilerlenir ve aynı zamanda proje için bir yol haritası belirlenmiş olur. CRISP-DM metodolojisi, Daimler Chrysler AG, SPSS, NCR ve OHRA gibi lider veri madenciliği kullanıcıları ve tedarikçilerinden oluşan bir konsorsiyum tarafından geliştirilmiş ve dünyanın en büyük veri madenciliği çözümlerinin kullandığı altı adımdan oluşmuş bir süreçtir. Bu altı adım aşağıda Şekil 1'de tanımlanmıştır.

Bu altı adım, daha detaylı olarak aşağıda verilmiştir.

Adım1: İş Hedeflerinin Belirlenmesi

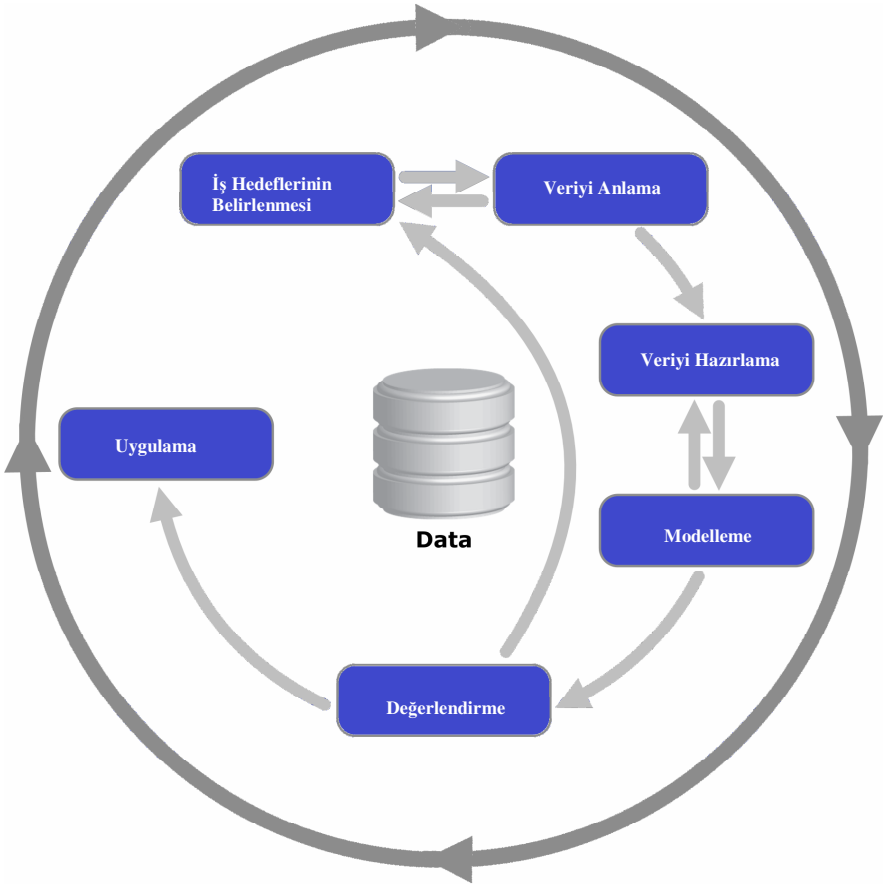
İlk adım proje amaçlarını ve iş gereksinimlerini anlamak, sonrasında da bu bilgiyi veri madenciliği problemi tanımına ve amaçlara ulaşmak için oluşturulan ilk proje planına dönüştürme üzerine odaklanır. Veri madenciliği projesindeki en zor kısım olan bu adımda, ne tür bir analiz yapılması gerektiğinin kesinlikle iyi bir şekilde anlaşılması gerekir. Aksi takdirde tüm proje yanlışlıklar üzerine kurulmuş olacak ve bulunan sonuçlar amacı temsil etmeyecektir. Bu yüzden ilk ve en önemli adım, amacı açıkça belirlemek ve amaca giden süreci geliştirmektir. Amacı tanımlarken, neyi ölçmeye veya öngörmeye çalıştığımıza karar vermek gerekir.

Adım 2: Veriyi Anlama

Veriyi anlama adımı öncelikle veriyi toplamakla başlar ve veri kümesinin içinde hangi değişkenlerin olduğunun saptamak, bu değişkenlerin ve değerlerinin neleri ifade ettiklerini anlamakla devam eder. Eğer analist veriyi tanımıyor ise, projenin diğer aşamalarına geçmeden önce verilere hakim bir kişiden yardım alıp veriyi anlamalıdır. Aksi takdirde, yanlış bir model oluşturmak söz konusu olabilir.

Adım 3: Veri Hazırlığı

Veri hazırlığı aşaması, ham veriden veri madenciliği aracında kullanılacak en son veri kümesini (mining data) oluşturmak için yapılan tüm işlemleri kapsamaktadır. Veri madenciliğinin en önemli aşamalarından bir tanesi olan verinin hazırlanması aşaması analistin toplam zaman ve enerjisinin %70 - %80'ini harcamasına neden olmaktadır. Bu aşamadaki görevlerden bazıları; tablo oluşturma, kayıt ve değişken seçimi, veri temizliği, yeni değişkenler oluşturma ve modelleme araçları için verileri dönüştürme (transformation) işlemleridir.



Şekil 1. CRISP-DM Döngüsü.

Adım 3: Veri Hazırlığı

Veri hazırlığı aşaması, ham veriden veri madenciliği aracında kullanılacak en son veri kümesini (mining data) oluşturmak için yapılan tüm işlemleri kapsamaktadır. Veri madenciliğinin en önemli aşamalarından bir tanesi olan verinin hazırlanması aşaması analistin toplam zaman ve enerjisinin %70 - %80'ini harcamasına neden olmaktadır. Bu aşamadaki görevlerden bazıları; tablo oluşturma, kayıt ve değişken seçimi, veri temizliği, yeni değişkenler oluşturma ve modelleme araçları için verileri dönüştürme (transformation) işlemleridir.

Adım 4: Modelleme

Bu adımda çeşitli modelleme teknikleri seçilip, uygulanır ve model parametreleri en uygun değerlere ayarlanır. Aynı tip veri madenciliği problemleri için çeşitli teknikler mevcuttur. Bazı teknikler belli veri formatlarına ihtiyaç duymaktadır. Bu yüzden genellikle veri hazırlama adımına geri dönüş gerekir. Klasik teknikler kullanıldığında tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar yinelenen bir süreçtir.

Model kuruluş süreci "Supervised" ve "Unsupervised" öğrenimin kullanıldığı modellere göre farklılık göstermektedir. Örnekten öğrenme olarak da isimlendirilen supervised öğrenimde, bir denetçi tarafından ilgili sınıflar önceden belirlenen bir kritere göre ayrılarak, her sınıf için çeşitli örnekler verilmektedir. Sistemin amacı verilen örneklerden hareket ederek her bir sınıfa ilişkin özelliklerin bulunması ve bu özelliklerin kural cümleleri ile ifade edilmesidir. Öğrenme süreci tamamlandığında, tanımlanan kural

cümleleri verilen yeni örneklere uygulanır ve yeni örneklerin hangi sınıfa ait olduğu kurulan model tarafından belirlenir. Denetimli öğrenme olarak da adlandırılan unsupervised öğrenmede ise, kümeleme analizinde olduğu gibi ilgili örneklerin gözlenmesi ve bu örneklerin özellikleri arasındaki benzerliklerden hareket ederek sınıfların tanımlanması amaçlanmaktadır.

Adım 5: Değerlendirme

Projenin bu aşamasında analistin elinde kalitesi yüksek bir model mevcuttur. Modelin uygulama aşamasına geçmeden önce modelin eksiksiz olarak değerlendirilmesi ve iş amaçlarına uyup uymadığına emin olmak için model oluşturulana kadar yürütülen adımların tekrar gözden geçirilmesi büyük önem taşımaktadır. Buradaki temel amaç, analiz süresince gözden kaçan önemli bir noktanın var olup olmadığını belirlemektir. Supervised öğrenimde seçilen algoritmaya uygun olarak ilgili veri kümesi hazırlandıktan sonra, ilk aşamada verinin bir kısmı modelin öğrenimi, diğer kısmı ise modelin geçerliliğinin test edilmesi için ayrılır. Modelin öğrenimi öğrenim kümesi kullanılarak gerçekleştirildikten sonra, test kümesi ile modelin doğruluk derecesi belirlenmektedir.

Adım 6: Uygulama

En iyi modeli oluşturmak genellikle veri madenciliği projesinin sonu değildir. Elde edilen bilginin düzenlenmesi ve müşterinin kullanacağı bir şekilde ifade edilmesi gerekmektedir. İhtiyaçlara göre uygulama adımı, bir rapor üretimi kadar basit veya oluşturulan modelin başka sistemlerin içine entegre edilmesi kadar karmaşık olabilir. Birçok durumda, uygulama adımlarını gerçekleştirecek kişi veri analisti değil, kullanıcı olmaktadır. Kurulan modeller risk analizi, kredi değerlendirme, dolandırıcılık tespiti gibi işletme uygulamalarında doğrudan kullanılabileceği gibi, promosyon planlaması simülasyonuna entegre edilebilir veya tahmin edilen envanter düzeyleri yeniden sipariş noktasının altına düştüğünde, otomatik olarak sipariş verilmesini sağlayacak bir uygulamanın içine gömülebilir.

Adım 7: İzleme ve Güncelleme

Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla ürettikleri veride değişiklikler ortaya çıkacaktır. Böyle bir durum karşısında modelin güncellenmesi gerekecektir. Günümüzde model güncelleme işleminin uzunca bir zaman alması ve efor gerektirmesi nedeniyle, işletmelerin çoğu bu tarz bir çalışma yapmamakta ve oluşturulmuş modelleri uzun yıllar boyunca kullanmaktadır.

Sahtecilik tespit projesi için oluşturulan bir model düşünüldüğünde, bu modelin 1 yıl boyunca hatta 2 ay boyunca bile sürekli kullanılmaması gerektiği bir gerçektir. Bunun sebebi, sahtekarların yakalandıkça taktiklerini değiştirmeye başlayacak olmalarıdır. Dolayısıyla yeni taktiklerle gelen yeni sahtekarların profilleri, oluşturulan modelde bulanan sahtekar profiline uymayacak ve model bu sahtekarları yakalayamaz duruma gelecektir. Böyle bir durum ile karşılaşmamak için sahtecilik modellerinin belirli aralıklarla güncellenmesi gerekmektedir.

Sonuç olarak güncelliğini yitirmiş modellerin tespiti için bu modellerin sürekli olarak izlenmesi gerekmektedir. Tahmin edilen ve gözlenen değişkenler arasındaki farklılığı gösteren grafikler model sonuçlarının izlenmesinde kullanılan yararlı bir yöntemdir [6].

2.2. Sınıflama Modelleri

Veri madenciliğinde modeller, tanımlayıcı ve tahmine dayalı modeller olmak üzere temel olarak ikiye ayrılır. Tanımlayıcı modellerde amaç, veri kümesindeki değişkenler arasındaki ilişkileri özetleyen örüntüleri ortaya çıkarmakken, tahmine dayalı modellerde belirli bir değişkenin değerini diğer değişkenlerden yararlanarak tahmin etmek amaçlanır [15].

Sınıflama modelleri, bağımlı değişkenin kategorik olduğu tahmine dayalı modellerdir. Bu modellerde bağımlı değişken "sınıf değişkeni" olarak da adlandırılabilir. Sınıflama modellerinde amaç, sınıflar

arasındaki sınırların belirlenerek bağımsız değişkenin verilen değerlerine göre bağımlı değişken değerini doğru bir şekilde tahmin etmektir. Literatürde en çok kullanılan sınıflama modellerine örnek olarak, karar ağaçları, lojistik regresyon, Bayesci ağlar, sinir ağları, en yakın komşuluk sınıflayıcıları, destek vektör makineleri verilebilir [15]. Bu çalışmada, bankacılık sektöründe doğrudan pazarlama verisinin çözümlenmesinde karar ağaçları, lojistik regresyon, Bayesci ağlar ve destek vektör makineleri gibi sınıflama yöntemleri kullanılmıştır. Bu yöntemler kısaca aşağıda tanıtılmıştır.

Lojistik Regresyon Modeli

Bağımlı değişkenin sürekli olduğu ve normal dağılım gösterdiği durumda bağımlı değişken ile bağımsız değişkenler arasındaki ilişkinin ortaya çıkartılmasında doğrusal regresyon modeli kullanılır. Ancak bağımlı değişken kategorik ise doğrusal regresyon modeli uygulanamaz. Bu durumda genelleştirilmiş doğrusal modeller ailesinin bir üyesi olan lojistik regresyon modeli kullanılabilir. İki düzeyli lojistik regresyon modeli iki farklı biçimde yazılabilir (Eşt.1, Eşt.2).

$$\log\left[\frac{P(y=1)}{1-P(y=1)}\right]=\log\left[\frac{P(y=1)}{P(y=0)}\right]=\sum_{j=1}^{k}\beta_j x_j \tag{1}$$

$$P(y=1)=\frac{\exp\left(\sum_{j=1}^{k}\beta_j x_j\right)}{1+\exp\left(\sum_{j=1}^{k}\beta_j x_j\right)} \tag{2}$$

Burada, y, yalnızca 0 ve 1 değerlerini alan bağımlı değişkeni; x_j, j=1,…,k olmak üzere j. bağımlı değişkeni, β_j ise j. bağımsız değişkene ait model parametresini vermektedir. Eşt.1'deki $P(y=1)/P(y=0)$ oranı ise bağımlı değişkenin tanımladığı olayın gerçekleşme olasılığının gerçekleşmeme olasılığına oranı olarak tanımlanan odds oranıdır.

Lojistik regresyon modelinde katsayı kestirimlerini elde etmek için En Çok Olabilirlik (EÇO), Ağırlıklı İteratif En Küçük Kareler (AİEKK) ve Minimum Lojit Ki-Kare (ML-χ^2) gibi yöntemler kullanılmaktadır [1].

Karar Ağaçları

Karar Ağaçları, adından da anlaşılabileceği gibi ağaç görünümünde tahmin edici bir yöntemdir. Karar ağacı yöntemini kullanarak verinin sınıflanması iki basamaklı bir işlemdir. İlk basamak öğrenme basamağıdır. Öğrenme basamağında önceden bilinen bir eğitim verisi, model oluşturmak amacı ile sınıflama algoritması tarafından çözümlenir. Öğrenilen model, sınıflama kuralları veya karar ağacı olarak gösterilir. İkinci basamak ise sınıflama basamağıdır. Sınıflama basamağında test verisi, sınıflama kurallarının veya karar ağacının doğruluğunu belirlemek amacıyla kullanılır. Eğer doğruluk kabul edilebilir oranda ise, kurallar yeni verilerin sınıflandırılmasında kullanılır. Karar ağaçlarının anlaşılması ve değiştirilmesi kolaydır. Karar ağaçları ile oluşturulan model, bir 'karar kuralları' kümesi içerir. Bu kurallar kümesi sınıf değişkeninin tahmin edilmesinde kullanılır.

Karar ağaçları, veri madenciliğinde en çok tercih edilen modellerden birisidir çünkü diğer yöntemlere oranla daha hızlı bir şekilde oluşturulabilir. Bu modellerin diğer bir üstünlüğü, basit bir yapıda olmaları ve anlaşılmalarının kolay olmasıdır [8].

Bayesci Ağlar

İnanç ağları olarak da bilinen Bayesci Ağlar, olasılıksal grafik modelleri ailesinin bir üyesidir. Olasılıksal grafik modelleri, ilgilenilen problemin kesin olmayan tanım kümesi hakkındaki bilgiyi temsil etmek için kullanılır. Bu grafiklerde düğümler (nodes) raslantı değişkenlerini, düğümler arasındaki bağlar (edges) ise raslantı değişkenleri arasındaki olasılıksal bağımlılık durumlarını gösterir. Bu bağımlılıklar genellikle bilinen istatistiksel ve sayısal yöntemlerden yararlanılarak analiz edilir. Bayesci ağlar, istatistik, makine öğrenmesi ve yapay zeka alanlarında çok kullanılan ve "yönlü dönüşsüz grafik (directed acyclic graph [DAG])" olarak bilinen bir grafiksel model yapısına sahiptir. Sezgisel olarak anlaşılabilir bir yapıya sahip olan bu ağlar, bir raslantı değişkenleri kümesinin çok değişkenli olasılık dağılımının etkili bir gösteriminin ve bu gösterim üzerinden çeşitli hesaplamaların yapılmasını sağlar [2].

Bayesci ağ yapısından yararlanarak, $V=\{X_1,\ldots,X_n\}$ değişken kümesi için tek bir çok değişkenli olasılık dağılımı tanımlanır ve bu dağılım Eşt. 3'den yararlanılarak elde edilir. Çok değişkenli olasılık dağılımının bu eşitlikten elde edilmesi "zincir kuralı (chain rule)" olarak adlandırılır [3, 9]

$$P(X_1,\ldots,X_n)=\prod_{i=1}^{n}P(X_i|\pi_i)=\prod_{i=1}^{n}\theta_{X_i|\pi_i} \tag{3}$$

Bir Bayesci ağ sınıflayıcısı, C sınıf değişkeni, $X_1,\ldots,X_n$ diğer değişkenler olmak üzere $V=\{X_1,\ldots,X_n,C\}$ değişken kümesi için oluşturulan bir Bayesci ağdır. $P(X_1,\ldots,X_n,C)$ olasılığını modelleyen bir Bayesci ağ B ile gösterilsin. Bu sınıflayıcı, B modeli üzerinden $P_B(c|\mathbf{x}_1,\ldots,\mathbf{x}_n)$ sonsal olasılığını maksimum yapan c sınıflarını belirler (Eşt. 4).

$$\arg\max_{C} P_B(C|X_1,X_2,\ldots,X_n) \tag{4}$$

Literatürde, sınıflandırma için kullanılan Bayesci ağ yapısını belirleyen birçok farklı yaklaşım vardır. Bunlardan bazıları, Naive Bayes, Tree Augmented Naive Bayes (TAN), BN Augmented Naive Bayes (BAN), Bayesian Multinet, Genel Bayesci ağ (GBN) sınıflayıcıları olarak sıralanabilir [5].

Destek Vektör Makineleri (DVM)

DVM, ikili sınıflama için kullanılan yeni öğrenme yöntemlerinden biridir. Bu yöntemin altındaki temel düşünce, d-boyutlu veriyi iki sınıfa ayıran bir hiper düzlem bulmaktır. Bununla birlikte, uygulamada veri kümesini doğrusal olarak ayırmak genellikle mümkün olmadığından çekirdek fonksiyonlardan yararlanılarak veri kümesi daha yüksek boyutlu uzaya bölünebilir. Bu durum, bazı hesaplama problemlerinin ve aşırı uyum sonucunun ortaya çıkmasına neden olur. Bu problemi ortadan kaldırmak için bu yüksek boyutlu uzayla doğrudan ilgilenmek yerine yalnızca bu uzaydaki nokta çarpım formülü ile ilgilenmek yeterli olacaktır. Ayrıca, sinir ağları gibi diğer sınıflama yöntemlerinden farklı olarak DVM'nin VC-boyutu açık bir şekilde hesaplanabilir. Sonuç olarak DVM, sezgisel, sağlam bir teorik dayanağı olan ve uygulamada başarılı bir sınıflama yöntemidir [4].

3. Doğrudan Pazarlama Stratejilerinin Belirlenmesinde Veri Madenciliği Süreci

Doğrudan pazarlamada veri madenciliğinin kullanımı ile ürünün/hizmetin olası alıcılarının belirlenmesi ve kitlesel pazarlamaya göre karın yükseltilmesi hedeflenir. Bankaların ve sigorta şirketlerinin, ürünlerini (krediler, emeklilik ve hayat sigortaları gibi) pazarlamak istedikleri müşterilerin oluşturduğu büyük veri tabanları vardır. Bu veri tabanlarında, doğrudan pazarlama için veri madenciliği problemleri iki durumda incelenebilir. Birinci durumda, veri tabanındaki müşterilerin % A'sı ürünü daha önce kitlesel pazarlama ya da pasif tanıtım yoluyla satın almıştır. A, genellikle 1 civarında değer alan küçük bir sayıdır. Burada veri madenciliği, veri tabanının % A'sını oluşturan ürünü satın almış olan müşterilere ilişkin örüntüleri

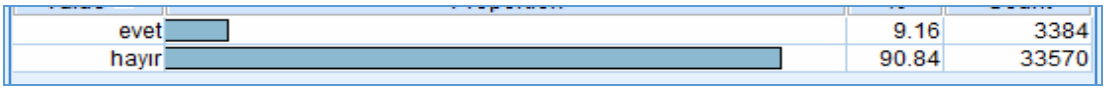
bularak; veri tabanının % (100-A)'sını oluşturan ve ürünü daha önce satın almayan kişilerden potansiyel müşterileri belirlemek için kullanılır.

İkinci durumda, piyasaya çıkan yeni bir ürün/hizmetin tanıtımı veri tabanındaki müşterilerin hepsine yapılır. Ürün/hizmet yeni olduğu için, veri tabanındaki müşterilerin hiçbiri daha önce satın almamıştır. Bu durumda, veri tabanındaki müşteriler arasından küçük bir kısmının (% 5 gibi) rasgele olarak seçildiği ve ürün/hizmet tanıtımı için hedef kitle olarak belirlendiği bir pilot çalışma yapılır. Bu pilot grubun yine %A'sının ürünü/hizmeti satın aldığı varsayılır. Birinci duruma benzer şekilde, bu pilot grup içerisinde potansiyel müşteriler belirlenmeye çalışılır.

Veri madenciliği süreci açısından her iki durumda gerçekleştirilen işlemler hemen hemen aynıdır. Tek fark, ikinci durumda daha küçük bir veri kümesi üzerinden hesaplamaların gerçekleştirilmesidir. Veri madenciliğinde kullanılan veri kümeleri genellikle çok büyük olduğundan bu durum bir sorun teşkil etmez [11].

Doğrudan pazarlama kampanyalarına ait veri kümelerinin veri madenciliği ile analizinde karşılaşılan önemli bir sorun "dengesizlik" olarak verilebilir. Bağımlı değişkenin kategorik olduğu durumlarda, kategoriler arasındaki dengesizlik sonuçları genellikle olumsuz yönde etkilemektedir. Daha önce bahsedildiği gibi, bu kampanyalara geri dönüş oranı (bağımlı değişkene olumlu cevap verenlerin oranı) genellikle % 1 gibi küçük bir sayıdır, geri kalan % 99'luk kısım ise negatif cevap vermektedir. Dolayısıyla dengesizlik, bağımlı değişkene ait dağılım modelleme yöntemlerinin sonuçlarını doğrudan etkileyen bir faktördür. Bu sorunun çözümü için çeşitli modelleme yöntemleri önerilmiştir. Yanlış sınıflama maliyeti, bağımlı değişkenin dengelenmesi, DVM gibi yöntemler bu sorunun çözümü için literatürde önerilen yöntemlerden bazılarıdır. Bu çalışma çerçevesinde, bağımlı değişkenin dengelenmesi yöntemi kullanılarak dengesizlik sorunu çözülmeye çalışılmış ve ele alınan yöntemlerin tamamı bu mantık çerçevesinde değerlendirilmiştir.

Bu çalışmada, Çizelge 1 ile tanımlanan veri kümesindeki bağımlı değişken olan "kampanya sonucu" için dağılım, %9 ve %91'dir. %9 kampanyaya geri dönüş yapan, %91 ise kampanyaya geri dönüş yapmayan müşterilerin oranını göstermektedir (Şekil 2).

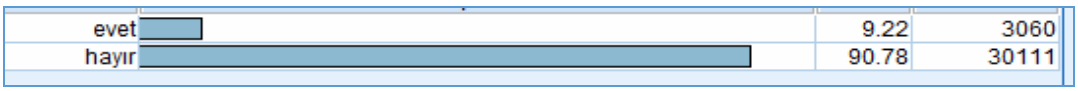


Şekil 2. Bağımlı değişkenin (kampanya sonucu) tüm veri kümesindeki dağılımı.

Uygulama bölümünde, dengesizlik sorununu çözmek için kullanılan yöntemde aşağıdaki süreç izlenmiştir:

i. Veri kümesi, %90'ı eğitim ve %10'u test veri kümeleri olmak üzere ikiye ayrılmıştır,

ii. Elde edilen eğitim veri kümesinde aşağıdaki dağılımlar elde edilmiştir,



Şekil 3. Bağımlı değişkenin (kampanya sonucu) eğitim veri kümesindeki dağılımı.

iii. "Evet" cevabını veren 3060 kayıt sabit tutularak, "hayır" cevabını veren 30111 kayıt içerisinden rasgele 3060 kayıt seçilerek bir örneklem oluşturulmuş ve elde edilen yeni veri kümesi ile modelleme yöntemleri kullanılmıştır.

iv. Sınıflama yöntemleri içerisinde literatürde sıklıkla kullanılan yedi algoritma, iii. maddesinde belirtilen mantık çerçevesinde rasgele olarak üçer kere denenmiştir.

v. Her bir yöntem için elde edilen üç model sonucu "güven-ağırlıklı oylama (confidence-weighted voting)" yöntemi ile tek bir sonuç haline getirilmiştir.

vi. Sonuçlar elde edilerek yorumlanmıştır.

4. Uygulama

Bu çalışmada, Portekiz'de Mayıs 2008 – Kasım 2010 tarihleri arasında gerçekleştirilen bir doğrudan pazarlama kampanyasında 45211 kişiden elde edilen bilgiye ait veri kümesinden yararlanılmıştır. Çalışmada kullanılan 16 bağımsız ve bir bağımlı değişken Çizelge 1'de tanımlanmıştır [12].

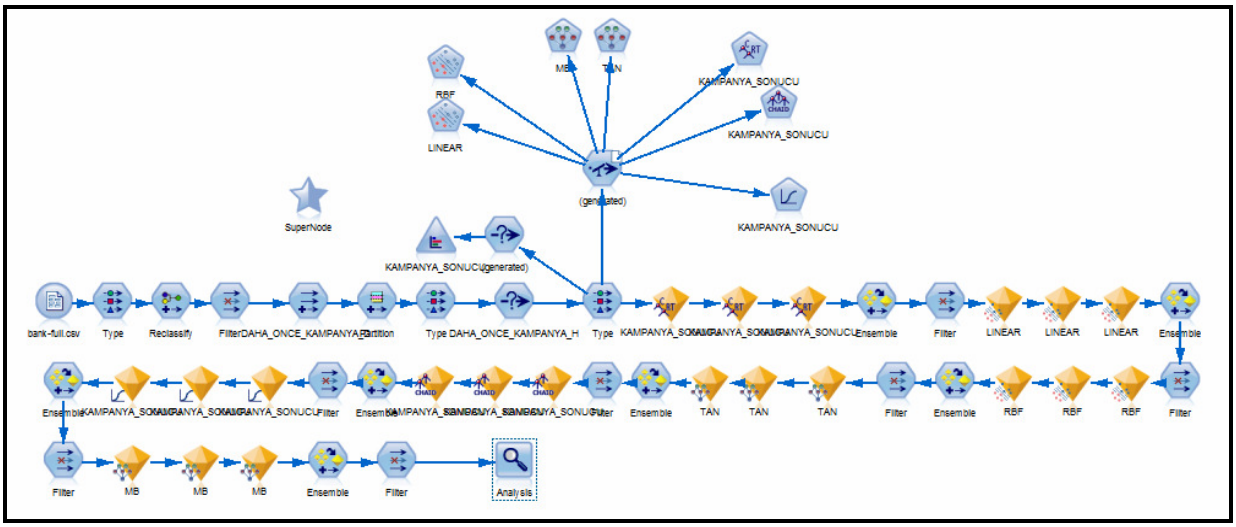
Çizelge 1. Doğrudan pazarlama kampanyasında yararlanılan değişkenler.

Bağımsız Değişkenler	**Açıklama**	**Düzeyleri**
age	Yaş	Sürekli
job	Meslek	Kategorik (yönetici, bilinmiyor, işsiz, işletmeci, ev hanımı, iş adamı, öğrenci, mavi yakalı, serbest meslek, emekli, teknisyen, hizmet işçisi)
marital	Medeni durum	Kategorik (evli, boşanmış-dul, bekar)
education	Eğitim durumu	Kategorik (bilinmiyor, ilköğretim, orta öğretim, yükseköğretim)
default	Kredi borcu ödememe	Kategorik (evet, hayır)
balance	Yıllık ortalama bakiye (euro)	Sürekli
housing	Üzerinde konut kredisi var mı?	Kategorik (evet, hayır)
loan	Üzerinde bireysel kredi var mı?	Kategorik (evet, hayır)
contact	İletişim kanalı	Kategorik (bilinmiyor, sabit telefon, cep telefonu)
day	Son iletişime geçilen gün (ayın kaçıncı günü?)	Sürekli
month	Son iletişime geçilen ay (yılın hangi ayı?)	Kategorik (Ocak, Şubat, Mart, …, Aralık)
duration	Son iletişim süresi (saniye)	Sürekli
campaign	İletişim kurma sayısı (son iletişim dahil)	Sürekli
pdays	Müşterinin bir önceki kampanya için son aranmasıyla şimdiki kampanya için ilk aranması arasında geçen süre (gün)	Sürekli
previous	Bu kampanyadan önce iletişim kurma sayısı	Sürekli
poutcome	Bir önceki kampanyanın sonucu	Kategorik (bilinmiyor, diğer, başarısız, başarılı)
Bağımlı Değişken	**Açıklama**	**Düzeyleri**
y	Kampanya sonucu (Müşteri vadeli mevduat hesabı açtırdı mı?)	Kategorik (evet, hayır)

3. Bölümde anlatılan dengesizlik sorunu ve bu sorunun çözümü için önerilen yöntem ışığında karar ağaçlarından, C&R Tree (Classification and Regression Tree), CHAID (Chi-squared Automatic Interaction Detection); istatistiksel yöntemlerden, lojistik regresyon; Bayesci ağlardan, TAN (Tree

Augmented Naïve Bayes), MB (Markov Blanket); destek vektör makinalarından, RBF (Radial Basis Function) ve Linear yöntemleri olmak üzere 7 farklı yöntem veri madenciliği çözümü olan IBM SPSS Modeler 15 programı yardımıyla karşılaştırılarak en iyi modelin hangisi olduğu tespit edilmiştir.

Şekil 4 ile verilen ekran görüntüsü bu karşılaştırmaları ve yöntemlere ilişkin model sonuçlarını (elmas) göstermektedir. Şekil 4'te görüldüğü gibi her model için üç sonuç (sarı elmas) elde edilmiştir.



Şekil 4. Veri kümesine uygulanan yöntemlere ve bu yöntemlerin karşılaştırılmasına ilişkin ekran görüntüsü.

Bağımlı değişkenin iki durumlu kategorik olduğu durumlarda sıklıkla kullanılan model başarı ölçütleri genel başarı (over-all), duyarlılık (sensitivity), seçicilik (specificity), F-Ölçütü (F-Measure), Matthews korelasyon katsayısıdır. Tüm modeller için elde edilen ölçüt değerleri Çizelge 2'de özetlenmiştir.

Çizelge 2. Doğrudan pazarlama kampanyasında yararlanılan değişkenler.

Yöntemler/Ölçüt	Genel Başarı	Duyarlılık	Seçicilik	F-Ölçütü	Matthews Korelasyon
C&R Tree	0,7275	0,9829	0,2212	0,8274	0,3447
CHAID	0,7891	0,9791	0,2644	0,8720	0,3845
LR	0,8123	0,9754	0,2836	0,8882	0,3930
TAN	0,6889	0,9510	0,1596	0,8035	0,1860
MB	0,6680	0,9520	0,1538	0,7869	0,1810
SVM Linear	0,8173	0,9746	0,2884	0,8173	0,3950
SVM RBF	0,8010	0,9747	0,2701	0,8801	0,3770

Yukarıda bahsedilen yöntemlerin hem tanımlayıcı hem de tahmin edici açıdan birbirlerine göre avantaj ve dezavantajları olmasına rağmen bu çalışmada sadece modellerin tahmin edici güçleri birbirleri ile karşılaştırılmıştır.

Yukarıda verilmiş olan tablodaki model başarı ölçütlerini yorumlamak gerekirse;

- Genel başarı ölçütü dikkate alındığında sırasıyla,
    - SVM Linear, lojistik regresyon ve SVM RBF yöntemleri en başarılı;
    - MB ve TAN ise en başarısız

tahmin yapan yöntemler olarak tespit edilmiştir.

- F ölçütü dikkate alındığında sırasıyla,
    - Lojistik regresyon, SVM RBF ve CHAID yöntemleri en başarılı;
    - MB ve TAN ise en başarısız

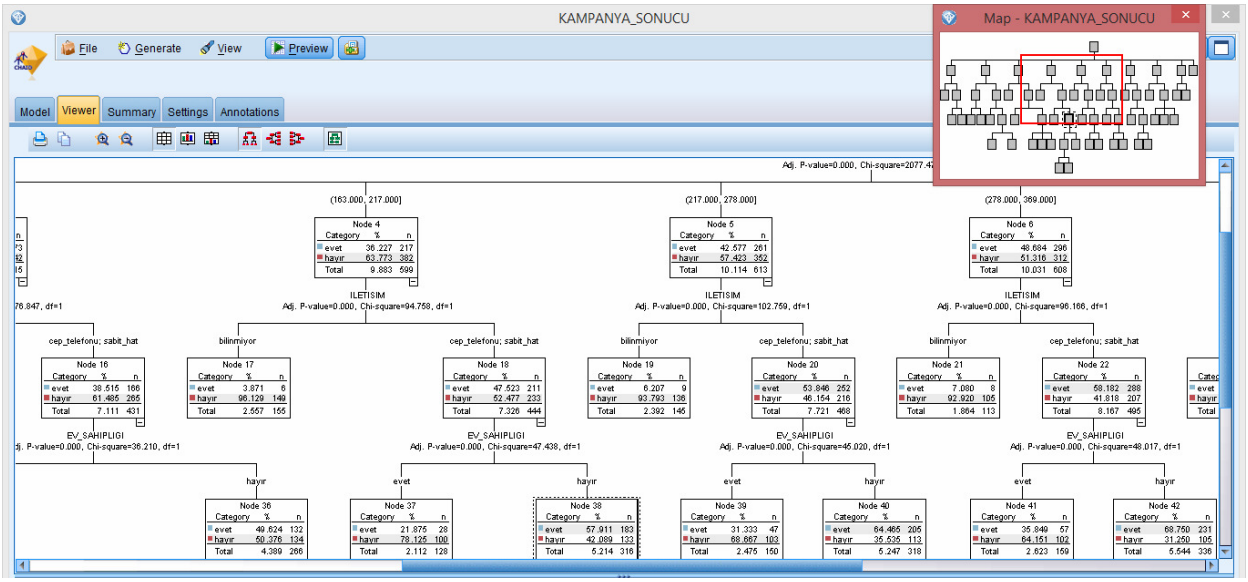
  tahmin yapan yöntemler olarak tespit edilmiştir.

- Matthews korelasyon katsayısı dikkate alındığında sırasıyla,
    - SVM Linear, lojistik regresyon ve CHAID yöntemleri en başarılı;
    - MB ve TAN ise en başarısız

  tahmin yapan yöntemler olarak tespit edilmiştir.

Dikkat edildiğinde, duyarlılık değerlerinin çok yüksek, seçicilik değerlerinin çok düşük olduğu göze çarpmaktadır. Bu durum, bağımlı değişkene ait prevelans değerinin dengesiz bir dağılıma sahip olmasından kaynaklanmaktadır. Böyle bir dengesizlik durumunda model başarısının test edilmesinde önerilen ölçütler F-Ölçütü ve Matthews korelasyon katsayısıdır. Matthews korelasyon katsayısı dikkate alındığında, lojistik regresyon, SVM Linear ve CHAID yöntemlerinin başarılı sonuçlar ürettiği tespit edilmiştir.

CHAID algoritmasının sonuçlarını daha detaylı yorumlamak için IBM SPSS Modeler 15 ile elde edilen karar ağacının çıktısı Şekil 5 ile verilmiştir.



Şekil 5. CHAID algoritması ile elde edilen karar ağacı çıktısı.

Buna göre;

i. "Kampanya sonucu" bağımlı değişkeni için en önemli değişken "son iletişim süresi" olarak tespit edilmiştir,

ii. "en son iletişim süresi" 163 – 217 saniye arasında olup, "iletişim kanalı" cep telefonu ve sabit olan ve

a. Üzerinde konut kredisi olmayan kişilerin %57.91 olasılıkla kampanyaya olumlu geri dönüş yapması beklenirken

b. Üzerinde konut kredisi olan kişilerin %78.12 olasılıkla kampanyaya olumsuz geri dönüş yapması beklenmektedir.

iii. "en son iletişim süresi" 670 – 941 saniye arasında olup, "iletişim kanalı" cep telefonu ise bu kişilerin %92.88 olasılıkla kampanyaya olumlu geri dönüş yapması beklenmektedir.

iv. Kampanyaya olumlu geri dönüş yapma olasılığı, "en son iletişim süresi" arttıkça artmaktadır.

CHAID algoritmasının sonucuna ilişkin sadece bazı genel sonuçlar verilmiştir. CHAID algoritması benzer 60 yorum daha üretmiştir.

Elde edilen sonuçlar ışığında, sınıflama algoritmalarının kullandığı bağımlı değişkenlere ait tablo Çizelge 3'te özetlenmiştir. Önemli olan bağımsız değişkenler "x" işareti ile işaretlenmiştir. "na" ile ifade edilen algoritmalar, bütün değişkenleri kullanarak işlem yapmaktadır.

Çizelge 3. Sınıflama algoritmalarına göre önemli bağımsız değişkenler.

	CHAID	CRT	LR	TAN	MB	SVM_LINEAR	SVM_RBF
age	x			na	x	na	na
job		x	x	na	x	na	na
marital	x		x	na	x	na	na
education	x			na	x	na	na
default				na		na	na
balance	x	x	x	na		na	na
housing	x	x	x	na	x	na	na
loan		x	x	na	x	na	na
contact	x	x	x	na	x	na	na
duration	x	x	x	na	x	na	na
campaign	x	x	x	na		na	na

5. Sonuç ve Tartışma

Bu çalışmada, bankacılık sektöründe doğrudan pazarlama kampanyaları için strateji belirlenmesinde veri madenciliği sürecinin uygulanması üzerinde durulmuştur. Bunun için CRISP-DM sürecinin adımları izlenmiş ve bu sürecin modelleme adımında bazı sınıflama algoritmalarından yararlanılmıştır. Ayrıca, doğrudan pazarlama kampanyalarının analizinde sıklıkla karşılaşılan bir problem olan bağımlı değişkenin kategorileri arasındaki dengesizlik problemi de göz önüne alınarak modellemeler gerçekleştirilmiştir. Portekiz'de gerçekleştirilen bir doğrusal pazarlama kampanyasına ilişkin veri kümesinde lojistik regresyon, karar ağaçları, Bayesci ağlar ve destek vektör makineleri sınıflama yöntemlerine ilişkin toplamda yedi farklı algoritma uygulanarak en iyi sonucu veren modeller farklı ölçümlere göre karşılaştırılmıştır. Dengesizlik durumu söz konusu olduğunda tercih edilen F ölçümü ve Matthews korelasyon katsayısı göz önüne alındığında lojistik regresyon ve karar ağaçlarından CHAID algoritmasının başarılı olduğu görülmüştür. CHAID algoritmasına göre bağımlı değişken "kampanya" sonucu için en önemli değişken "son iletişim süresi" olarak belirlenmiştir. Ayrıca, "iletişim kanalı" "üzerinde konut kredisi var mı?" değişkenlerinin de bağımlı değişken üzerinde etkili olduğu görülmüştür. Tüm algoritmalar için önemli bulunan değişkenler incelendiğinde ise benzer şekilde "son iletişim süresi", "iletişim kanalı" ve "üzerinde konut kredisi var mı?" değişkenlerinin tüm algoritmalarda önemli bulunmuştur. Sonuç olarak, bankacılık sektöründeki doğrudan pazarlama kampanyalarının bu üç

değişkene daha çok ağırlık verilerek hazırlanmasının kampanyanın olumlu sonuç oranını yükseltmesi açısından önemli olduğu söylenebilir.

Kaynaklar

[1] A. Agresti, *Categorical Data Analysis*, Wiley, 744p.

[2] I. Ben-Gal, 2007, *Bayesian Networks*, Encyclopedia of Statistics in Quality &Reliability,F. Ruggeri, F.Faltin, R. Kenett, R. (eds), Wiley & Sons.

[3] S.G. Boettcher and C. Dethlefsen, 2003, Deal: A package for learning Bayesian networks, *Journal of Statistical Software*, 8(20), 1-40.

[4] D. Boswell, 2002, Introduction to Support Vector Machines, http://www.work.caltech.edu/~boswell/IntroToSVM.pdf.

[5] J. Cheng and R. Greiner, 2001, Learning Bayesian Belief Network Classifiers: Algorithm and System, *Proceedings of the Fourteenth Canadian Conference on Artificial Intelligence (AI'2001).*

[6] M.Ö. Dolgun, 2014, *Veri Madenciliği Sınıflama Yöntemlerinin Başarılarının; Bağımlı Değişken Prevelansı, Örneklem Büyüklüğü ve Bağımsız Değişkenler Arası İlişki Yapısına Göre Karşılaştırılması,* Doktora Tezi, Sağlık Bilimleri Enstitüsü, Hacettepe Üniversitesi.

[7] D. Hand, H. Mannila, P. Smyth, 2001, *Principles of Data Mining*, The MIT Press, Cambridge, 546p.

[8] IBM, 2012, IBM DB2 version 9.5 manual, http://publib.boulder.ibm.com/infocenter/db2luw/v9r5/index.jsp?topic=%2Fcom.ibm.datatools.datamining.doc%2Fc_decision_tree_calssification.html.

[9] F.V. Jensen, 2001, *Bayesian Networks and Decision Graphs*, Springer-Verlag, New York, 268p.

[10] D.T. Larose, 2004, *Discovering Knowledge in Data: An Introduction to Data Mining*, Wiley Interscience, New York, 222p.

[11] C.X. Ling and C. Li, 1998, Data mining for direct marketing: problems and solutions, *Proceedings of KDD'98*, 217-225.

[12] S. Moro, R.M:S: Laureano, P. Cortez, 2000, Using data mining for bank direct marketing: an application of the CRISP-DM methodology, *Proceedings of the European Simulation and Modelling Conference (ESM'2011)*, Guimares, Portugal, October 2011, s.117-121.

[13] P.V. Putten, 1999, Data mining in direct ;marketing databases, *Complexity and Management: A Collection of Essay,* World Scientific.

[14] E. Simoudis , B. Livezey, R. Kerber, 1996, *Integrating inductive and deductive reasoning for data mining.* Advances in Knowledge Discovery and Data Mining. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy,R. (eds), The MIT Press, Cambridge, s.353-374.

[15] P. Tan, M. Steinbach, V. Kumar, 2006, *Introduction to Data Mining*, Addison-Wesley, Boston, 769p.