Doğrudan Pazarlama Faaliyetlerinde Yatırım Kararının Tahmininde Sınıflandırma Algoritmalarının Karşılaştırılması

Nur Kuban Torun[1] **Tolga Torun**[2]

[1]Bilecik Şeyh Edebali Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, İşletme Bölümü

[2]Bilecik Şeyh Edebali Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Yönetim Bilişim Sistemleri Bölümü

Özet: Günümüz rekabet şartlarında işletmeler için gerçek hedef kitlesine ulaşarak bu hedef kitlelerinden maksimum geri dönüşü elde etmeleri, başarı için önemli bir unsurdur. Bu yöntemlerden bir tanesi doğrudan pazarlama faaliyetleridir. Özellikle bankacılık sektöründe doğrudan pazarlama müşteriye ulaşmada kullanılan etkin bir yöntemdir. Doğrudan pazarlama tekniği içerisinde yer alan telemarketing ile bankaların müşteriye yatırım yapma ya da yapmama kararı konusunda en iyi tahminde bulunması açısından teknikler karşılaştırılmıştır. Bu teknikler lojistik regresyon ve Naive Bayes yöntemleridir. Uygulanan yöntemler literatür taramasından elde edilen 17 müşteri değişkeni göz önünde bulundurularak kendi içerisinde değerlendirme tabi tutulmuş ve en etkin yöntem olarak lojistik regresyon yöntemi bulunmuştur.

Anahtar Kelimeler: Doğrudan Pazarlama, Lojistik Regresyon, Naive Bayes.

JEL Kodları: M39; C35; C81

The Comparison of Classification Algorithms in Estimation of Investment Decisions in Direct Marketing Activities

Abstract: Nowadays, it is important to determine literal target consumers achieving success with maximum feedback in competitive market places. One of the efficient technicals of detecting target consumers is direct marketing. Especially telemarketing technique in direct marketing is the best way to investigate consumers investing decisions. In this study, logistic regression and Naive Bayes techniques are compared and data are obtained from telemarketing. Due to the literature view 17 kind of consumer variables are determined and these variables are analyzed. As a result of analyze logistic regression has found the efficient technique.

Keywords: Direct Marketing, Logistic Regression, Naive Bayes.

JEL Codes: M39; C35; C81

1. Giriş

Pazarlama içerisinde, müşterilerden elde edilen gelire baktığımızda, bu gelirin %80'nini eski müşterilerinin (Cook ve Mindak, 1984), geri kalan kısmını yani %20'sini ise yeni müşterilerin oluşturduğunu görmekteyiz (Stewart, 1995). Bu açıdan değerlendirildiğinde, pazarlama faaliyetleri içerisinde mevcut yani eski müşterilere doğrudan ulaşılmasının önemi ortaya çıkmaktadır (Kaefer, vd. 2005). Bu anlayış doğrultusunda doğrudan pazarlama, müşterilerden hayat boyu değer elde edilerek, kâr elde edilmesi için önemli bir faktördür (Wang, vd. 2005).

Doğrudan pazarlama, Amerikan Doğrudan Pazarlama Birliği (U.S. Direct Marketing Association) "*kayda değer etki yaratmak ya da satın alma davranışı oluşturmak amacıyla, her hangi bir pazarlama aracını kullanarak bir ya da bir çok pazarlama reklamının kullanıldığı etkileşimli pazarlama faaliyeti*" olarak tanımlamaktadır. Bu tanımdan yola çıkarak doğrudan pazarlamanın ilk olarak etkileşime açık bir yapısı olduğunu görmekteyiz. Bu açıdan doğrudan pazarlama, işletmeler ve müşteriler açısından iki yönlü bir iletişimin kapısını açmaktadır. Doğrudan pazarlamanın, kitlesel pazarlamada yer alan tek yönlü monologlarından farklı olarak, müşterilerin doğrudan pazarlama iletişimi unsurlarına cevap vermelerine bağlı olarak, müşteriler ile ilgili değerli bilgileri işletmelere sunan ve müşterilere yönelik bir kişisellik içeren yapısı bulunmaktadır. Diğer yandan doğrudan pazarlama faaliyetleri içerisinde yer alan iletişim araçları sayesinde, pazarlama faaliyetleri coğrafik engellerin üstesinden gelmektedir. Son olarak ise doğrudan pazarlama ölçülebilir ve sayılabilir bir yapıya sahiptir. Bu

açıdan doğrudan pazarlama faaliyetleri ile müşterilerin satın alma ya da almama oranlarının bilgisine rahatlıkla ulaşılabilmektedir (Des ve Lee, 1994).

Kitlesel pazarlama içerisinde işletmeler, pazarlama faaliyetlerinin kim için iyi; kim için kötü olduğunun kararını verememektedir. Kitlesel pazarlama faaliyetleri, potansiyel olsun ya da olmasın tüm tüketiciler için yapılmakta ve gerçek manada müşteriler için uygun olan uygulamalar kullanılamamaktadır. Ancak doğru müşteriye ulaşmak özellikle ürün/hizmet ile ilgili tanıtımlar, numuneleri tanıtmak ya da doğru hedef kitleye doğru ürünleri sunmak açısından önem arz etmektedir. Bu açıdan kitlesel pazarlama faaliyetleri oldukça maliyetli olmasına rağmen geri dönüşleri oldukça düşüktür (Kaefer, vd. 2005). Kitlesel pazarlama faaliyetlerinin tam zıttı olarak ortaya çıkan doğrudan pazarlama faaliyetleri ile en iyi uygulamaları potansiyel ve doğru müşterilere sunma imkanı yaratılmaktadır (Kaefer, 2005). Bu açıdan değerlendirildiğinde doğrudan pazarlama faaliyetleri, karar verme aşamasında gerekli olan önemli bilgiyi sağlamakta ve bu sayede zaman tasarrufu sağlamaktadır. Ancak doğrudan pazarlamaya konu olacak hedef kitlenin iyi bir şekilde incelenmesi gerekmektedir. Günümüzde bu elde edilen yüksek hacimdeki büyük bilgilerin (big data) analizinde kullanılan bir çok analiz tekniği bulunmaktadır (Moro, vd. 2012).

Doğrudan pazarlama özellikle finans alanında yer alan bankacılık sektöründe oldukça sık kullanılan bir uygulama olarak yerini almış bulunmaktadır. Bankacılık sektörü içerisinde arz ve talep bakımında oldukça büyük bir akış sağlanmaktadır. Bu akışa bağlı olarak da rekabet oldukça yoğundur (Des ve Lee, 1994). Özellikle Internet, dijital pazarlama, telefonun yaygınlaşması, mobil telefon ve SMS'lerin gelişmesine bağlı olarak bankacılık sektörünün müşterilere ulaşmada doğrudan pazarlama faaliyetlerini kullanmaya yönelik ilgisi artmış bulunmaktadır (Hagel ve Rayport, 1997). Internet'in yaygınlaşmasına rağmen, bankacılık sektöründe doğrudan pazarlama faaliyetleri kapsamında, telefon bankacılığı halen cazibesini kaybetmemiş durumdadır. Bu doğrultuda bankacılık sektöründe, telemarketing (telefonda pazarlama) şeklinde doğrudan pazarlama faaliyetlerine halen önem verilmektedir. Telemarketing ile bankalar müşterilerine ürün/hizmet hakkında bilgi vermekte ve satış yapmaktadır. Telemarketing sırasında müşterilerin bilgilerine ulaşılması, doğrudan pazarlamanın etkinliği için oldukça önemli bir faktördür. İster telemarketing isterse diğer doğrudan pazarlama iletişimi araçları vasıtasıyla elde edilmiş müşterileri veri bankaları, onlarla iletişimin kurulmasında, bilgilerin analiz edilmesinde ve bu bilgiler doğrultusunda reklamların çeşitlendirilmesinde önemli bir unsurdur (Vajiramedhin ve Suebsing, 2014).

Doğrudan pazarlama iletişimi araçlarından yararlanılırken yoğun bir şekilde müşterilerin yaş, cinsiyet, gelir gibi demografik bilgileri veri olarak kullanılmakta; ancak bu faktörler hedef kitleyi etkilemede yetersiz olmaktadır (Gıpta ve Chintagunta, 1994). Araştırmalar sadece demografik faktörlerin değil; aynı zamanda geçmiş deneyimlerin de inceleme altına alınması gerektiğini ortaya koymaktadır. Bu sayede işletmeler, doğrudan pazarlama faaliyetleri içerisinde hedef kitlesine tam olarak uygun uygulamaları ortaya koyabilmektedir (Guadagni ve Little, 1983; Heilman, vd. 2000). Gelişen teknoloji ile birlikte pazarlamacıların karşısında tek bir kriterden daha çok analiz edilmesi gereken milyonlarca kriter bulunmaktadır (Robertshaw ve Marr, 2005). Doğrudan pazarlama faaliyetlerinde en büyük güçlük, bu değişkenlerin bir formüle uyarlanmasında yatmaktadır. Bu formülasyonun içinde, müşterilerin gelecek kampanya aralığında işletmeyi tercih edip etmeyeceği de yer almaktadır. Bu açıdan uygulanacak kampanyanın hangi aralıkta olacağının belirlenmesi özellikle belirli bir bütçeye sahip müşteriler göz önünde bulundurularak yapılması gerekmektedir (Baesens, vd. 2002).

Diğer yandan, doğrudan pazarlama faaliyetleri uygulamacılar açısından kısa vadeli uygulamalar ile desteklenmekte ve uzun vadede oluşacak müşteri beklentilerinde ve satın alma davranışlarındaki değişmeler göz ardı edilmektedir. Uygulamacılar, zaman içerisinde müşterilerin işletme ile etkileşimini göz ardı etmekte ve uzun vadede kârlılık sağlanamamaktadır (Pednault, vd. 2002). Bu sayılan olumsuzluklardan kurtulmanın yolu müşterilerin işletme ile ilişkisini bitirme olasılıklarını göz önünde bulundurmak ve bu olasılıklar dahilinde faaliyetlerini sürdürmesi gerekmektedir. Bu açıdan müşterilerin hangi hallerde işletme ile değişim ilişkilerine devam ettiklerinin ortaya koyulması önemli bir faktördür (Kim, vd. 2009).

Doğrudan pazarlama faaliyetlerinde bulunulurken, müşterilerin mevcut konumlarının belirli değerler açısından sınıflandırılması gerekmektedir.

Tablo 1: Veri setinin değişkenlerinin tanımı

Değişken İsimleri	Açıklamaları	Segment	Türleri
Age(yaş)	Müşterinin yaşı		Nümerik
Job(iş)	Müşterinin çalıştığı iş türü	Lider(admin) Girişimci Mavi yakalı Emekli Teknisyen Öğrenci Yönetici Serbest meslek Hizmet sektörü Bilinmiyor Hizmetçi İşsiz	Kategorik
Marital(medeni durum)	Medeni durum	Evli Boşanmış(dul veya ayrılmış) Bekar Diğer	Kategorik
Education(eğitim)	Eğitim durumu	Orta okul İlk okul Lise	Kategorik
Default(mevcut durum)	Şu an hali hazırda aldığınız kredi var mı?	Evet Hayır	Kategorik
Balance(bakiye)	Ortalama yıllık bakiye (euro üzerinden)		Nümerik
Housing(konut)	Konut borcu var mı?	Evet Hayır	Kategorik
Loan(borç)	Kişisel borcu var mı?	Evet Hayır	Kategorik
Contact(iletişim şekli)	İletişim şekli	Diğer Ev telefonu Telefon	Nümerik
Day(gün)	En son iletişime geçilen aydaki gün		Nümerik
Month(ay)	En son iletişime geçilen ay		Nümerik
Duration(iletişim süresi)	En son arandığında kaç saniye konuşma yapıldı?		Nümerik
Campaign(kampanya)	Bu kampanya için kaç kez aynı müşteri ile iletişime geçildi?		Nümerik
Pdays (geçen günler)	Bir önceki kampanya ile ilgili iletişime geçildikten sonra geçen gün sayısı		Nümerik
Previous (önceki)	Mevcut kampanyadan önce kaç kez müşteri ile iletişime geçildi		Nümerik
Poutcome(geçmişe ait getiri)	Daha önceki kampanyalardan getiri durumu	Diğer Bilinmiyor Başarısız Başarılı	Kategorik
y	Müşteri yatırım yapmayı kabul etti mi?	Evet Hayır	Kategorik

Sınıflandırma özellikle veri madenciliğinin temelini oluşturmaktadır. Sınıflandırma yapılırken tıbbi tanılar, örnek tanımlar, endüstri içerisindeki temel hatalar, finansal pazar içerisindeki ortak eğilim gibi etkenler göz önünde bulundurulmaktadır. Her bir sınıflandırmaya tabi olacak kalem, yoğun bir öğrenmenin ve geri bildirimin neticesidir. Sınıflandırma yapan kişiler, bir sınıflandırma modeli oluşturmakta ve tahmin için gerekli olan konuları bu sınıflandırmaya dahil etmektedir. Bu sınıflandırma içerisinde yer alan kalemlere girdi denilmektedir. Bu model içerisinde yer alan kalemler, if-then (eğer-sonra) kurallarından, karar ağaçlarından ya da nöral ağlarlardan türetilmiştir (Vajiramedhin ve Suebsing, 2014).

2. Verinin Elde Edilmesi

Veri Portekiz bankalarının doğrudan pazarlama kampanyaları ile ilgilidir. Verinin kaynağı The UCI Machine Learning Repository'den elde edilen gerçek verilere dayanmaktadır. Veri seti Portekiz tedarikçi bankalarından elde edilmiş ve Mayıs 2008 ile Kasım 2013 yılları arasında yapılmış 45.212 telefon konuşmasına dayanmaktadır. Evet ya da hayır gibi seçenekleri seçmesine bağlı olarak, bazı müşteriler ile birden fazla kez iletişime geçildiği olmuştur. Veri seti Tablo 1'de yer alan 17 adet değişken içermektedir.

3. Metodoloji

3.1. Lojistik Regresyon

Lojistik regresyon birden fazla türde konuyu içeren veri setlerinin analizi için uygun ve etkili bir metot olarak karşımıza çıkmaktadır. Buna ek olarak, kategorik değişkenleri ve bir ya da birden fazla kategorik veya sürekli bağımlı tahmin edici olduğunda kullanılan bir tekniktir (Elsalamony, 2014).

Lojistik regresyon ayrıca gelenekler çoklu regresyon yöntemlerinde kullanılan en küçük kareler tekniği yerine maksimum olasılık tahminlerini kullanılmaktadır. Başlangıç değeri olarak tahmin edilen parametreleri kullanmaktadır ve örneklemin olasılığı bu parametrelerin hesaplandığı popülasyondan gelmektedir. Büyük olasılık değeri elde edilene kadar tahmin edilen parametre değerleri yinelenerek uyarlanır. Böylelikle, maksimum olasılık yaklaşımı, en çok istenilen gözlemlenen veriyi bulmaya çalışmaktadır (Kleinbaum ve Klein, 2010).

Ancak lojistik regresyon yaklaşımı şu formun öğrenen fonksiyonu şeklindedir: F: A →Y, veya P(Y/A), Y'nin hedef olarak ayrık değerli olduğu durumda, ve A= (A_1, A_2,...,A_n) ayrık içerek bağımlı, flag veya sürekli bağımsız nitelik iken. Bankalar ile ilgili bu çalışmada, doğrudan pazarlama veri seti içerisinde Y flag niteliktir (evet ve hayır) ve seçilen bölünmüş veri içinde ileri binom süreci seçeneği bulunmaktadır (Elsalamony, 2014).

Lojistik formül Y=1 (veya evet) olasılığı üzerine kuruludur ve bu P ile sembolize edilmektedir. Y'nin 0 olduğu (veya hayır) olasılık 1-P'dir. Bu denklem (1):

$$\ln\left(\frac{P}{1-P}\right) = w_0 + w_1 A$$

Burada w_0+w_1A regresyon denkleminden aşina olduğumuz denklemdir. Lojistik regresyon P(Y\A) dağılımının parametrik formunu ön görmektedir. Buradan yola çıkarak ele alınan veri seti içerisindeki parametreleri doğrudan tahmin etmektedir. Parametrik model (2) ve (3):

$$P(Y = yes|A) = \frac{\exp\left(w_0 + \sum_{i=1}^{n} w_i A_i\right)}{1 + \exp\left(w_0 + \sum_{i=1}^{n} w_i A_i\right)}$$

ve

$$P(Y = no|A) = \frac{1}{1 + \exp\left(w_0 + \sum_{i=1}^{n} w_i A_i\right)}$$

W_0 burada denklemin kısıtıdır. w_1 ise yordayıcı değişkenin katsayısıdır. Logits (log odds) olarak bilinen (1) ile gösterilen denklem, lojistik denklemin katsayısıdır (eğim değeri). Eğim, bir birim A'da yaşanan değişimin, Y'nin ortalama değerindeki değişikliği olarak düşünülebilir.

Lojistik regresyonun birçok faydası bulunmaktadır. Lojistik regresyonda bağımsız değişkenlerin normal dağılıma sahip olma veya her grupta eşit varyansa sahip olma, girdilerin ve bağımlı değişkenin doğrusal ilişkiye sahip olma şartı bulunmamaktadır. Ayrıca varyansların homojenliğine ve hata terimlerinin normal dağılımına bakılmaksızın açık etkileşimler ve kuvvet terimleri eklenebilmektedir. Bağımsız değişkenlerin aralıklı veya sınırsız olması da gerekmemektedir. Bu sayılan avantajların yanında lojistik regresyon oldukça çok veriye sahip olma koşulunu beraberinde getirmektedir. Geleneksel regresyon modellerinde her bir değişken için 20 veri yeterli iken; lojistik regresyonda bu sayı 50'dir.

3.2. Naïve Bayes Sınıflandırma

Naïve Bayes (TAN) sınıflandırma en etkili ve etkin bir sınıflandırma algoritmasıdır. Bayesyan ağların özel bir vakasını içermektedir. Sınırlandırılmamış bayesyan ağların yapısı ve parametreleri iyileştirmeler için anlamlı sonuçlar vermektedir. Tan, Friedman (1997) tarafından kıyaslama yapmaya yarayan büyük verilerden hızlı sonuç almak amacıyla geliştirilmiştir. Bayesyan sınıflandırma bir gruba veya sınıfa ait belirli bir parçayı tahmin etmekte kullanılmaktadır. Bu teknik, hızlı olmasından dolayı ve büyük veri setlerinde tutarlı sonuçlar alınmasından dolayı tercih edilmektedir. Dataları tahmin etme sırasında TAN, oldukça küçük bir eğitim veri setine ihtiyaç duymakta ve gerçek ve soyut verilerin sınıflandırılmasında başarı sağlamaktadır. Bayesyan sınıflandırmanın en büyük dezavantajı, bazı durumlarda ele aldığı sorunların gerçek problemler olmamasıdır. Bu teknik, sadece istatistik alanına değil; her hangi seçilen bir alanda rahatlıkla uygulanabilmektedir (Wikipedia, 2015).

Verilen bir x'in (x = [x(1), x(2), . . . , x(L)]^T ∈ R^L) sınıf S_i'ye ait olup olmadığına karar vermek için kullanılan yukarıda formüle edilen Bayes karar teoreminde istatistik olarak bağımsızlık önermesinden yararlanılırsa bu tip sınıflandırmaya Naive Bayes sınıflandırılması denir. Matematiksel bir ifadeyle:

$P(x|S_i)P(S_i) > P(x|S_j)P(S_j)_{,\forall j \neq, i}$ ifadesindeki $P(x|S_i)$ terimi yeniden aşağıdaki gibi yazılır:

$$P(x|S_i) \approx \prod_{k=1}^{L} P(x_k|S_i)$$

böylece Bayes karar teoremi aşağıdaki şekli alır. Bayes karar teorisine göre x sınıf S_i'ya aittir, eğer;

$$P(S_i) \prod_{k=1}^{L} P(x_k|S_i) > P(S_j) \prod_{k=1}^{L} P(x_k|S_j)$$

$P(S_i)_{ve}$ $P(S_j)$ i ve j sınıflarının öncel olasılıklarıdır. Elde olan veri kümesinden değerleri kolayca hesaplanabilir.

Naive Bayes sınıflandırıcının kullanım alanı her ne kadar kısıtlı gözükse de yüksek boyutlu uzayda ve yeterli sayıda veriyle x'in (nicelik kümesi) bileşenlerinin istatistik olarak bağımsız olması koşulu esnetilerek başarılı sonuçlar elde edilebilinir.

4. Kullanılan Modelin Ölçütleri

Modelin başarısını ölçmek için kullanılan temel kavramlar hata oranı, kesinlik, duyarlılık ve F-ölçütüdür. Modelin başarısı, doğru sınıfa atanan örnek sayısı ve yanlış sınıfa atılan örnek sayısı nicelikleriyle alakalıdır (Coşkun ve Baykal, 2011).

Doğruluk oranı: Doğru sınıflandırılmış örnek sayısının toplam örnek sayısına oranıdır.

$$doğruluk\ oranı = \frac{TP + TN}{TP + FP + FN + TN}$$

Hata oranı: Yanlış sınıflandırılmış örnek sayısının toplam örnek sayısına oranıdır.

$$hata\ oranı = \frac{FP + FN}{TP + FP + FN + TN}$$

Kesinlik oranı: Pozitif olarak tahmin edilen doğru örnek sayısının, pozitif olarak tahminlenen tüm örnek sayısına oranıdır.

$$kesinlik\ oranı = \frac{TP}{TP + FP}$$

Duyarlılık oranı: Doğru sınıflandırılmış pozitif örnek sayısının; toplam pozitif örnek sayısına oranıdır.

$$duyarlılık\ oranı = \frac{TP}{TP + FN}$$

Özgüllük(belirlilik oranı): doğru negatiflerin toplam yanlışlara oranıdır.

$$özgüllük = \frac{TN}{TN + FP}$$

$$F - ölçütü = \frac{2 * duyarlılık\ oranı * kesinlik\ oranı}{duyarlılık + kesinlik}$$

Tablo 2: Karışıklık Matrisi

Karışıklık matrisi		Gerçek	
		Sınıf=1	Sınıf=0
Tahmin	Sınıf=1	TP(doğru pozitif)	FP(yanlış pozitif)
	Sınıf=0	FN(yanlış negatif)	TN(doğru negatif)

5. Uygulama

Bu çalışmada, telemarketing şeklinde doğrudan pazarlama uygulamasının gelecek dönemlerde yapılacak yatırım kararlarlarını tahminlemek için sınıflandırma algoritmalarından naive bayes ve diğer bir yöntem olan lojistik regresyon yöntemleri kullanılmıştır. Yöntemler R programında çözülmüştür. Uygun paketler uygulamaya başlanmadan önce yüklenmiştir. Veri seti 45212 gözlem değerinden oluşmaktadır. Bunun %10 kısmı 4522 tanesi de test verisidir. 17 adet değişken bulunmaktadır. Veriler uygulamaya başlanılmadan uygun hale getirilmiştir. Bunun için binary kategorik verilere 1-0, kategorik verilere sayılar sıralı sayı değerleri ve nümerik verilerden yaş, bakiye, gün (en son iletişime geçilen aydaki gün), iletişim süresi (en son iletişim süresi), kampanya (bu kampanya için kaç kez aynı müşteri ile iletişime geçildi), geçen günler (bir önceki kampanya ile ilgili iletişime geçildikten sonra geçen gün sayısı), önceki (mevcut kampanyadan önce kaç kez müşteri ile iletişime geçildiğinin sayısı) değerlerine aralık belirlenmiştir. Daha sonra bu aralıklı verilerde kategoriğe dönüştürülmüş ve iki analizde bu şekilde çözümlenmiştir.

5.1. Naive Bayes Sınıflaması

Öncelikle veri setimizi R programında çözdüreceğimiz için naive bayes için uygun olan "e1071" adlı paket kurulmuştur.

Toplam 45212 verinin %10' nu 4522 tanesi test verisi olarak kullanılmıştır. Öncelikle nümerik veriler aralıklı verilere daha sonra da kategorik verilere dönüştürülmüştür. Uygulama yapılırken verilerin tamamı faktör olarak tanımlanarak çözdürülmüştür.

Test datamızda alınan sonuçlarda;

Modelde 3967 test verisinin değeri 0 olarak, 484 tanesinin değeri de 1 olarak tahminlenmiştir.

5.1. Lojistik regresyon

Lojistik regresyon yöntemini R da uygulamadan önce "aod" ve "ggplot2" ve "caret" paketlerini kuruyoruz. Toplam 45212 verimiz var Biz bunun %10' nu 4522 tanesi test verisi olarak kullandık.

Eğitim verisi ile modeli kurduğumuzda anlamlı çıkan değişkenler şu şekilde olmuştur. Anlamlılık düzeyi=000 olanlar burada gösterilmiştir. Ek bölümünde tablo 9'da detaylı şekilde incelenebilinir.

Yaş değişkeni için anlamlı: age2 [30-39] yaş aralığında olanlar, age3[40-49] yaş aralığında olanlar,age4 [50-59] yaş aralığında olanlar, age5 [60-69] yaş aralığında olanlar

İş değişkeni için anlamlı: job3(mavi yakalılar) ve job11 (hizmetçiler)

Medeni durum değişkeni için: m2 (bekar) ve m3 (boşanmış)

Eğitim değişkeni için: e3 (lise mezunu)

Konut borcu değişkeni için: h1(evet) yani konut borcu olanlar

Borç değişkeni için: l1(evet) kişisel borcu olanlar

İletişim şekli için: ct2 (telefon) ct3 (cep telefonu)

Gün değişkeni için: day5[12-14] günleri ve day11[30-31] günleri

Ay değişkeni için: month2(şubat), month3(mart), month4(nisan), month5(mayıs), month6(haziran), month8(ağustos), month9(eylül), month10(ekim), month11(kasım), month12(aralık)

İletişim süresi için: dr2[300-599], dr3[600-899], dr4[900-1199], dr5[1200-1499], dr6[1500-1799], dr7[1800-2099], dr8[2100-2399], dr9[2400-2699], dr10[2700-2999], dr11[3000-3299]

5.2. ROC Eğrisi

Testin ayırt etme gücünün belirlenmesinde kullandık. Tanı testi ne kadar iyiyse eğri o kadar yukarıya ve sola doğru kayar. (0,0)-(0-1),(1-1) yanlış değerlere sahip olmayan ideal bir testte rock eğrisidir. Buna karşın y=x doğrusuna çizim yaklaştıkça başarısız bir test ortaya çıkar (Dirican, 2001).

ROC puanı 0.8871Çıkmıştır. Bu puan ROC eğrisinin altında kalan oranı temsil eder. ROC eğrisine göre bu modelde doğrudan pazarlama ile yatırım yapmama durumunu, yatırım yapma durumundan ayırt etme başarısı %88'dir.

Tablo 3: Naive Bayes İkili Sınıflandırma Matrisi

		Gerçek	
		0	1
Tahmin	0	3738	299
	1	262	222

Tablo 4: Naive Bayes Sınıflandırma Algoritmaları Oranları

Naive- Bayes	**Test seti**
Doğruluk oranı	%87,5
Hata oranı	%12,4
Kesinlik oranı	%92,5
Duyarlılık oranı	%93,4
Özgüllük oranı	%42,6
F-Ölçüt	%93

Lojistik regresyon ikili matrisi şu şekildedir:

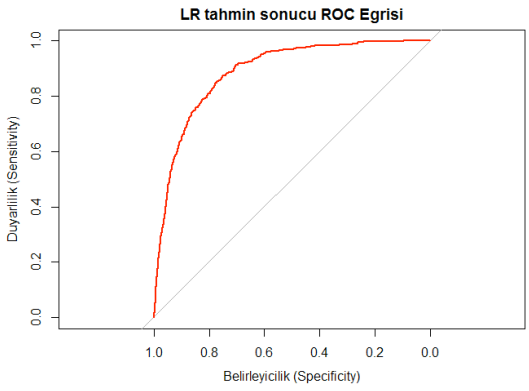
Tablo 5: Lojistik Regresyon İkili Sınıflandırma Matrisi

		Gerçek	
		0	1
Tahmin	0	3911	386
	1	89	135

4297 test verisinin değeri 0, 224 test verisinin değeri ise 1 çıkmıştır.

Tablo 6: Lojistik Regresyon Sınıflandırma Algoritmaları Oranları

Lojistik regresyon	Test seti
Doğruluk oranı	%89,4
Hata oranı	%10,5
Kesinlik oranı	%91
Duyarlılık oranı	%97,7
Özgüllük oranı	%25,9
F-ölçütü	%94,2



Şekil 1: Roc Eğrisi

Tablo 7: Oluşturulan modellerin başarım ölçütleri

	TP	FP	FN	TN	Doğruluk oranı	Hata oranı	Kesinlik oranı	Duyarlılık oranı	Özgünllük oranı	F-ölçütü
Naive-Bayes	3738	299	262	222	%87,5	%12,4	%92,5	%93,4	%42,6	%93
Lojistik regresyon	3911	386	89	135	%89,4	%10,5	%91	%97,7	%25,9	%94,2

5.3. Oluşturulan Modellerin Başarım Ölçütleri

Uygulanan 2 yöntem kıyaslandığında, lojistik regresyonda hem TP (doğru pozitif), hem FP (yanlış pozitif) hem de FN (yanlış negatif) değerleri daha yüksektir. Naive Bayes'te TN(doğru negatif) değerleri daha yüksektir. Doğruluk oranı, duyarlılık oranı ve F-ölçütü lojistik regresyon algoritmasında daha yüksek bir ona sahiptir. Hata oranı, kesinlik oranı ve özgünlük oranı değerleri ise Naive Bayes'te daha yüksektir.

6. Sonuçlar

Bankacılık sektörü içerisinde kullanılan doğrudan pazarlama faaliyetlerinde doğru hedef kitlenin belirlenmesi özellikle kampanyaların etkinliği ve müşteri kazanma açısından önemlidir. Bu açıdan tüketicilere ait verilerin iyi bir şekilde değerlendirilerek, potansiyel müşterilerden oluşan bir hedef kitlenin tahmin edilmesi, etkinliği arttıran bir faktör olarak karşımıza çıkmaktadır. Bu açıdan değerlendirildiğinde lojistik regresyon yöntemi potansiyel müşterileri en iyi şekilde tahmin edebilmektedir.

Lojistik regresyon ile yaş, iş niteliği, medeni durum, eğitim durumu, konut borcunun olup olmaması, borç değişkeni, iletişim şekli, en son aranılan gün ve ay değişkeni ve de toplam telemarketing yardımıyla yapılan konuşma süresi en etkili unsurlardandır.

Lojistik regresyonda hem TP (doğru pozitif), hem FP (yanlış pozitif) hem de FN (yanlış negatif) değerleri daha yüksektir. Naive Bayes'te TN(doğru negatif) değerleri daha yüksektir

Lojistik regresyonda doğruluk oranı %89.4 iken Naive Bayes'te %87.5'tir. Doğruluk oranına göre değerlendirildiğinde lojistik regresyonun daha iyi sonuç verdiğini görmekteyiz.

Duyarlılık oranlarına bakıldığında lojistik regresyonun %97.7, Naive Bayes'in ise %93.4 olduğunu görmekteyiz. Duyarlılık oranlarında yine lojistik regresyonun en iyi sonucu verdiğini görmekteyiz.

Kesinlik ölçütü dikkate alındığında ise %92.5 oran ile Naive Bayes iyi sonuç vermiştir. Lojistik regresyonun kesinlik ölçütü %91 olarak bulunmuştur.

Kesinlik ve duyarlılık ölçütlerinin harmonik ortalaması olan F-ölçütüne bakılarak bu iki ölçüt bir arada değerlendirilmiştir. Bu ölçüte göre lojistik regresyonun %94.2 oran ile en iyi sonucu verdiğini görmekteyiz. Diğer yandan Naive Bayes'in F-ölçüt değeri %93'tür.

Rakamlar doğrultusunda özetlersek doğruluk oranı, duyarlılık oranı ve F-ölçütü lojistik regresyon algoritmasında daha yüksek bir ona sahiptir. Hata oranı, kesinlik oranı ve özgünlük oranı değerleri ise Naive Bayes'te daha yüksektir.

Tüm bu çıkarımlardan sonra, lojistik regresyon algoritması, diğer algoritmaya göre daha başarılı bulunmuştur.

Kaynaklar

Baesens, B, Viaene, S, Van Den Poel, D, Vanthienen, J ve Dedene, G. (2002). Bayesian Neural Network Learning for Repeat Purchase Modeling in Direct Marketing. European Journal of Operational Research, 138(1), 191-211.

Cook, V.J. ve Mindak, W.A. (1984). A Search for the Constants: The 'Heavy User' Revisited! Journal of Consumer Marketing, 1, 79–81.

Coşkun, C. Ve Baykal, A. (2011). Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması. Akademik Bilişim'11 - XIII. Akademik Bilişim Konferansı Bildirileri, 2 - 4 Şubat 2011 İnönü Üniversitesi, Malatya: 51-58.

Des, T. ve Lee, S. C. I. (1994). Direct Marketing in the Financial Services Industry. Journal of Marketing Management, 10: 377-390.

Dirican, A. (2001). Tanı Testi Performanslarının Değerlendirilmesi ve Kıyaslanması. Cerrahpaşa J Med, 32(1): 25-30.

Guadagni, P.M., ve Little J.D.C. (1983). A Logit Model of Brand Choice Calibrated on Scanner Data. Marketing Science, 2: 203–238.

Gupta, S. ve Chintagunta, P.K. (1994). On Using Demographic Variables to Determine Segment

Membership in Logit Mixture Models. Journal of Marketing Research, 31: 128–36.

Hagel, J. ve Rayport, J. F. (1997). The Coming Battle for Customer Information. Harvard Business Review, January-February, 53.

Heilman, C.M., Bowman D. ve Wright G.P. (2000). The Evolution Of Brand Preferences and Choice Behaviors of Consumers in a New Market. Journal of Marketing Research, 37, 139–155.

Kaefer, F., Heilman, C. ve Ramenofsky, S. (2005). A Neural Network Application to Consumer Classification to Improve the Timing of Direct Marketing Activities. Computers & Operations Research, 32(10), 2595-2615.

Kim, Y; Song, H and Kim, S (2009). A New Marketing Strategy Map for Direct Marketing. Knowledge-Based Systems, 22(5), 327-335.

Kleinbaum D.G. ve Klein, M. (2010). Logistic Regression, Statistics for Biology and Health. Springer Science Business Media, LLC.

Moro, S., M. S. R. Laureano ve P. Cortez (2012). Enhancing Bank Direct Marketing Through Data Mining. CAlg European Marketing Academ: 1–9.

Pednault, E., Abe, N. ve Zadrozny, B. (2002). Sequential Cost-Sensitive Decision Making With Reinforcement Learning. in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD '02), ACM Press.

Popelka, O., Hrebicek, J., Stenel, M., Hodinka, M. ve Trenz, O. (2012). Comparison of Different Non-Statistical Classification Methods. 30th International Conference Mathematical Methods in Economics: 727-732.

Robertshaw, G ve Marr, N (2006). An Empirical Measure of the Availability, Completeness and Reliability of Voluntarily Disclosed Personal Information for Direct Marketing Purposes. Journal of Financial Services Marketing, 11(1): 85-94.

Stewart T.A. (1995). After All You've Done for Your Customers, Why Are They Still Not Happy? Fortune, 11: 178–82.

UCI (2015). https://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

Vajiramedhin, C. ve Suebsing, A. (2014). Feature Selection with Data Balancing for Prediction of Bank Telemarketing. Applied Mathematical Sciences, 8 (114): 5667-5672.

Wang, K., Zhou, S. ve Yeung J.M.S. (2005). Mining customer value: From association rules to direct marketing. Data Mining and Knowledge Discovery, 11: 57–79.

Wikipedia, (2015). http://en.wikipedia.org/wiki/Naïve_Bayes_classifier Wikipedia has a tool to generate citations for particular articles related to Naïve Bayes classifier.

EK:

Ek 1: Uygulamada Kullanılan Değişkenlerin Kategorik Değerleri

Age(yaş)	%	Job(iş)	%	Marital (medeni durum)	%	Education (eğitim)	%
[18-29]=1	%12	Admin=1	%11	Married=1	%60	Primary=1	%16
[30-39]=2	%40	Entrepreneur=2	%3	Single=2	%28	Secondary=2	%51
[40-49]=3	%25.4	Blue-collar=3	%22	Divorced=3	%12	Tertiary=3	%29
[50-59]=4	%18.6	Retired=4	%5			Unknown=4	%4
[60-69]=5	%2.7	Technician=5	%17				
[70-79]=6	%0.9	Student=6	%2				
[80-89]=7	%0.27	Management=7	%21				
[90-99]=8	%0.02	Self-employed=8	%3.7				
		Services=9	%9				
		Unknown=10	%0.64				
		Housemaid=11	%2.74				
		Unemployed=12	%2.88				

Default (mevcutd)	%	Balance (bakiye)	%	Housing (konut)	%	Loan (borç)	%
Yes=1	%2	[-8019-9999]=1	%98.17	Yes=1	%56	Yes=1	%16
No=0	%98	[10000-19999]=2	%1.4	No=0	%44	No=0	%84
		[20000-29999]=3	%0.31				
		[30000-39999]=4	%0.05				
		[40000-49999]=5	%0.02				
		[50000-59999]=6	%0.02				
		[60000-69999]=7	%0.008				
		[70000-79999]=8	%0.02				
		[80000-89999]=9	%0.009				
		[90000-99999]=10	%0.002				
		[100000-109999]=11	%0.002				

Contact (iletişim)	%	Day(gün)	%	Month(ay)	%	Duration (süre)	%
Unknown=1	%29	[0-2]=1	%4	Jan=1	%3	[0-299]=1	%73
Cellular=2	%65	[3-5]=2	%10	Feb=2	%6	[300-599]=2	%19
Telephone=3	%6	[6-8]=3	%12	Mar=3	%1	[600-899]=3	%5
		[9-11]=4	%8	Apr=4	%6	[900-1199]=4	%2
		[12-14]=5	%11	May=5	%30	[1200-1499]=5	%1
		[15-17]=6	%11	Jun=6	%12	[1500-1799]=6	%0
		[18-20]=7	%15	Jul=7	%15	[1800-2099]=7	%0
		[21-23]=8	%9	Aug=8	%14	[2100-2399]=8	%0
		[24-26]=9	%5	Sept=9	%1	[2400-2699]=9	%0
		[27-29]=10	%10	Oct=10	%2	[2700-2999]=10	%0
		[30-31]=11	%5	Nov=11	%9	[3000-3299]=11	%0
				Dec=12	%1	[3300-3599]=12	%0
						[3600-3899]=13	%0
						[3900-4999]=14	%0

Campaign (kampanya)	%	Pdays (geçen gün)	%	Previous (önceki)	%	Poutcome (geçmişgetiri)	%
[1-9]=1	%96.7	-1=1	%81.7	[0-24]	%99.93	unknown	%82
[10-19]=2	%2.6	[0-99]=2	%3.1	[25-49]	%0.06	success	%3
[20-29]=3	%0.43	[100-199]=3	%6.4	[50-74]	%0.01	failure	%11
[30-39]=4	%0.2	[200-299]=4	%3.3	[75-275]	%0.001	other	%4
[40-49]=5	%0.04	[300-399]=5	%5				
[50-59]=6	%0.03	[400-499]=6	%0.3				
[60-69]=7	%0.001	[500-599]=7	%0.1				
		[600-699]=8	%0.02				
		[700-799]=9	%0.06				
		[800-899]=10	%0.02				

Ek 2. Lojistik Regresyon Eğitim Setinden R Programı İle Alınan Kısıt ve Anlamlılık Değerleri

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-5.097835	0.198320	-25.705	< 2e-16	***
age2	-0.378556	0.059445	-6.368	1.91e-10	***
age3	-0.381030	0.068930	-5.528	3.24e-08	***
age4	-0.452063	0.076904	-5.878	4.15e-09	***
age5	0.484527	0.114279	4.240	2.24e-05	***
age6	0.443203	0.164875	2.688	0.007186	**
age7	0.434584	0.266830	1.629	0.103378	
age8	1.991674	0.938582	2.122	0.033837	*
job2	-0.323383	0.126420	-2.558	0.010528	*
job3	-0.264299	0.072846	-3.628	0.000285	***
job4	-0.165819	0.109622	-1.513	0.130371	
job5	-0.140338	0.069515	-2.019	0.043505	*
job6	0.271802	0.112488	2.416	0.015680	*
job7	-0.155359	0.074067	-2.098	0.035947	*
job8	-0.340831	0.112814	-3.021	0.002518	**
job9	-0.189035	0.084121	-2.247	0.024629	*
job10	-0.412934	0.234107	-1.764	0.077754	.
job11	-0.569394	0.136337	-4.176	2.96e-05	***
job12	-0.168690	0.112460	-1.500	0.133613	
m2	0.214466	0.046466	4.616	3.92e-06	***
m3	0.220893	0.059259	3.728	0.000193	***
e2	0.188298	0.065090	2.893	0.003817	**
e3	0.402451	0.075898	5.303	1.14e-07	***
e4	0.237158	0.105224	2.254	0.024206	*
d1	-0.056768	0.163006	-0.348	0.727644	
b2	0.155248	0.138204	1.123	0.261298	
b3	-0.084414	0.295570	-0.286	0.775186	
b4	-0.117894	0.697331	-0.169	0.865745	
b5	0.332867	1.180873	0.282	0.778033	
b6	0.627030	1.047219	0.599	0.549334	
b7	-11.138338	293.002331	-0.038	0.969676	
b8	-12.462501	535.411189	-0.023	0.981430	
b9	11.530679	372.898070	0.031	0.975332	

b10	-10.003675	535.411181	-0.019	0.985093
b11	-11.825375	535.411180	-0.022	0.982379
h1	-0.609366	0.044191	-13.789	< 2e-16 ***
l1	-0.370082	0.059320	-6.239	4.41e-10 ***
ct2	1.582726	0.071737	22.063	< 2e-16 ***
ct3	1.340383	0.100388	13.352	< 2e-16 ***
day2	-0.004753	0.100118	-0.047	0.962139
day3	-0.043999	0.105058	-0.419	0.675358
day4	0.221434	0.107222	2.065	0.038905 *
day5	0.371494	0.104971	3.539	0.000402 ***
day6	-0.001227	0.104294	-0.012	0.990613
day7	-0.201589	0.106008	-1.902	0.057219 .
day8	0.261472	0.110370	2.369	0.017834 *
day9	0.229427	0.117693	1.949	0.051252 .
day10	0.188355	0.114140	1.650	0.098900 .
day11	0.449064	0.122558	3.664	0.000248 ***
month2	1.110682	0.136769	8.121	4.63e-16 ***
month3	2.929769	0.158116	18.529	< 2e-16 ***
month4	1.379433	0.129296	10.669	< 2e-16 ***
month5	0.885510	0.126232	7.015	2.30e-12 ***
month6	1.769675	0.139642	12.673	< 2e-16 ***
month7	0.317553	0.124543	2.550	0.010780 *
month8	0.490081	0.125639	3.901	9.59e-05 ***
month9	2.008386	0.158084	12.705	< 2e-16 ***
month10	2.136834	0.147790	14.459	< 2e-16 ***
month11	0.601620	0.136102	4.420	9.85e-06 ***
month12	1.998489	0.206199	9.692	< 2e-16 ***
dr2	1.551453	0.043395	35.752	< 2e-16 ***
dr3	3.073089	0.057035	53.881	< 2e-16 ***
dr4	3.894981	0.084239	46.237	< 2e-16 ***
dr5	4.220581	0.128002	32.973	< 2e-16 ***
dr6	4.573526	0.218260	20.954	< 2e-16 ***
dr7	3.721017	0.279822	13.298	< 2e-16 ***
dr8	2.124714	0.637385	3.333	0.000858 ***
dr9	4.437787	0.677867	6.547	5.88e-11 ***
dr10	4.353214	1.189463	3.660	0.000252 ***
dr11	4.023730	0.877597	4.585	4.54e-06 ***
dr12	-9.919942	298.792546	-0.033	0.973515
dr13	4.056275	1.467156	2.765	0.005697 **
dr14	-9.832353	535.411188	-0.018	0.985348
cam2	-0.513773	0.158132	-3.249	0.001158 **
cam3	-1.032335	0.531802	-1.941	0.052234 .
cam4	-0.574430	1.016464	-0.565	0.571988
cam5	-9.352902	195.817491	-0.048	0.961905
cam6	-10.743226	202.472076	-0.053	0.957684
cam7	-8.243557	535.411186	-0.015	0.987716
pdays2	0.449249	1.006432	0.446	0.655324
pdays3	-0.073955	1.005343	-0.074	0.941359
pdays4	-0.365285	1.009414	-0.362	0.717443
pdays5	-0.631314	1.008506	-0.626	0.531322
pdays6	1.773984	1.018442	1.742	0.081533 .
pdays7	1.337922	1.034615	1.293	0.195956
pdays8	0.496249	1.187876	0.418	0.676121
pdays9	1.241518	1.169649	1.061	0.288488

```
pdays10      1.665275   1.221841   1.363 0.172907
pre2        -0.782088   0.795026  -0.984 0.325250
pre3         3.539549   1.345663   2.630 0.008530 **
pre4       -11.966700 535.411192  -0.022 0.982168
pt2          2.387005   1.005901   2.373 0.017644 *
pt3          0.210428   1.004804   0.209 0.834119
pt4          0.362196   1.006671   0.360 0.719000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```