Çok Yanıtlı Doğrusal Regresyon Algoritmasıyla Lojistik Regresyon ve Birliktelik Kuralı Sonuçlarının Birleştirilmesi

Nevzat Ekmekçi[1], Özge Yücel Kasap[2], Utku Görkem Ketenci[2], Oya Kalıpsız[1], Mehmet S. Aktaş[1]

[1] Bilgisayar Mühendisliği Bölümü, Elektrik-Elektronik Fakültesi
Yıldız Teknik Üniversitesi, İstanbul
[2] Ar-Ge Merkezi, Cybersoft, İstanbul
nvztekmekci@gmail.com, ozge.kasap@cs.com.tr, utku.ketenci@cs.com.tr,
kalipsiz@yildiz.edu.tr, aktas@yildiz.edu.tr

Özet Pazarlama alanının en önemli noktalarından biri, doğru ürünü, doğru müşteriye önermektir. Bu çalışma, yeni bir veri madenciliği aracı olarak geliştirilen PROPCA'nın bir parçası olarak bu soruna bir çözüm önermektedir. Bu çalışmanın amacı, lojistik regresyon ve birliktelik kurallarını bir araya getirip ürün önerileri yapmaktır. Bu algoritmaların ayrı ayrı kullanımlarından daha iyi sonuç veren bir yöntem olarak, lojistik regresyon ve birliktelik kurallarının birleşimi sunulmuştur. Birliktelik kuralları verilen veri kümesindeki tüm kuralları ararken, lojistik regresyon müşteriler için belirli bir ürünü satın alma olasılığı tahmininde bulunmaktadır. Bu iki yaklaşımın birleşimi bankacılık alanında yer alan gerçek bir veri kümesi üzerinde test edilmiş, sonuçlar karşılaştırılmış ve genel olarak uygunlukları tartışılmıştır.

Anahtar Kelimeler: Birliktelik Kuralları, Lojistik Regresyon Analizi, Veri Madenciliği, Toplu Öğrenme, İstifleme, Çok Yanıtlı Doğrusal Regresyon.

Özet One of the keys in marketing is to recommend the right products to the right customers. This paper proposes a solution to this problem as a part of the development of a new data mining tool PROPCA. The aim is to use logistic regression analysis and association rule mining together to make recommendations in marketing. An approach in which combination of these two algorithms provides better results than algorithms used standalone is presented. While association rule mining searches all rules in the data set, logistic regression predicts a purchase probability of a product for customers. The combination of these two approaches are tested on a real-life banking data set. The results of combination are shown and their suitability in general is discussed.

Anahtar Kelimeler: Association Rules, Logistic Regression, Data Mining, Ensemble Learning, Stacking, Multi-response Linear Regression.

1 Giriş

İnsanlar, geniş ürün yelpazesinden dolayı, doğru hizmet veya ürün seçiminde zorlanmaktadır. Diğer yandan pazarlama alanında, hangi ürünlerin hangi müşterilerin ilgisini çekeceğini ya da hangi müşterilerin ilgisini çekmeyeceğini bilmek önemlidir. Böylece, müşterinin ilgisiz olacağı ürünler hakkında kampanya postası veya mesaj alması engellenebilmektedir.

Bu durum, finans ve bankacılık sektöründe önemli bir sorundur. IBM Silverpop tarafından 2014 yılında Amerika'da finans sektöründe yapılan bir araştırmaya göre [1], bu e-postaların açılma ve tıklanma oranları sırasıyla %22.4 ve %3.3' dür. Bu istatistikler göz önüne alındığında, eğer her ürün için potansiyel alıcılar düzgün tespit edilirse, yanlış kişilere gönderilen mesaj, arama ve e-posta sayılarında azalma sağlanarak, şirketlerin karlılığının artması sağlanabilir. Geniş ürün yelpazesinin büyümesi göz önüne alındığında, bir pazarlama uzmanı için, müşteri davranışlarını anlama ve açıklamada, istatistiksel modellerin veya makine öğrenmesi algoritmalarının kullanılması kaçınılmazdır.

Son yıllarda, Lojistik Regresyon (LR) modelleri, birçok alanda tahminler yapmak için yaygın olarak kullanılmaktadır [2]. Daha belirli olması açısından; Akıncı, pazarlama araştırmalarında, LR modelinin uygulamalarını incelemiştir. Bunun sonucunda, tüketici davranışı modelleme, uluslararası pazarlama, toplumsal pazarlama, perakende promosyon ve sağlık hizmetleri pazarlaması gibi alanları ortaya koymuştur [3].

Benzer şekilde, Birliktelik Kuralları (BK) da Veri Madenciliği (VM)'nin pazarlama alanındaki önemli çalışmalarından biridir [4]. BK, sık sık birlikte satın alınan 2 veya daha fazla ürünü ortaya çıkarmayı amaçlamaktadır.

Bu çalışmanın asıl amacı, çok sayıda müşteri için, gerçekten satın alabilecekleri ürünleri önerebilecek, daha güvenilir bir model oluşturmaktır. Büyük pazarlama verileri göz önüne alındığında, VM algoritmalarının tüm veri kümesi üzerinde yeterli performans gösteremediği görülmüştür. Bunun yerine, veri setinden daha küçük bir örnek küme VM algoritmalarını uygulamak için seçilebilir ve sonuçları tüm müşterilerin yer aldığı büyük veri kümesi üzerinde genelleştirilebilir. Ancak, daha iyi doğruluk elde etmek için, VM algoritmalarını birlikte kullanmak kaçınılmazdır.

Bu araştırma, pazarlama alanında yaygın olarak kullanılan iki modeli; LR ve BK'nı, daha iyi sonuçlar elde etmek için birleştirmeyi önermekte ve bu yaklaşımı incelemektedir. Bu iki model, PROPCA üzerinde bulunmaktadır. PROPCA, aykırı değer analizi, eksik veri tamamlama, özellik seçimi, faktör analizi, örnekleme, lojistik regresyon, kümeleme analizi ve birliktelik algoritmaları yöntemlerini içeren yeni bir VM aracıdır. LR ve BK birbirlerinden oldukça farklı yollar izlemektedirler. LR tahmin yapmak için müşteri ve ürün özelliklerini kullanırken, BK sadece ürün ailelerinin sahiplik bilgilerini kullanmaktadır. Toplu Öğrenme (TÖ), bu iki tamamlayıcı modeli birleştirerek, tüketici davranışı modellemede yeni bir yaklaşım ortaya koyacaktır.

Bu bildirinin yazım organizasyonu şu şekildedir; 2. bölümde; temel kavramlar ve LR, BK, TÖ hakkında yer alan ilgili çalışmalar açıklanacaktır. 3. bölümde; önerilen model tanımlanmaktadır. Çalışmanın sonuçlarının yanı sıra, uygulama

ve testler 4. bölümde verilmektedir. Son olarak 5. bölümde; son hükümler ve gelecek araştırmalar için tavsiyeler ele alınacaktır.

2 İlgili Çalışmalar

Daha önce de belirtildiği gibi bu araştırmanın amacı, daha iyi tahmin (sınıflandırma) sonuçları elde etmek için, TÖ yöntemleriyle LR ve BK sonuçlarını birleştirmektir. Bu bölümde, ilk olarak BK hakkında yer alan çalışmalar tanıtılacaktır. Ardından, LR modeliyle ilgili çalışmalar ele alınacaktır. Sonuncu, fakat bir o kadar da önemli olarak, TÖ ile ilgili araştırmalar kısaca sunulacaktır.

2.1 Birliktelik Kuralları

BK, VM'nin en önemli alanlarından biridir [5]. BK, verideki en anlamlı ilişkiler kullanılarak bulunan Sık Rastlanan Öge Kümeleri (SRÖK) sonucunda oluşturulur. BK, müşteri davranışlarını analiz ve tahmin etmek için kullanılabilir [6]. Başlangıçta bu yöntem, hangi ürünün, hangi ürünle satın alınacağını bulmak için 'Pazar Sepet Analizi' alanında kabul edilmiştir. Verilen bir dizi müşteri satın alma davranışları (farklı müşteriler tarafından birlikte satın alınan ürünlerin kayıtları) göz önüne alındığında, BK'nın amacı bu işlemlerde en sık satın alınan ürünleri veya öğeleri bulmaktır. Diğer yandan, ürün kümelemesi ve katalog tasarımı için de yararlıdır.

Bir öğe kümesi veri tabanında yer alan tüm ürünlerin bir alt kümesidir. İşlemlerse, müşterilerin birlikte satın aldığı ürün gruplarıdır. $X \rightarrow Y$, BK için bir örnektir ve aynı zamanda X ve Y ürün gruplarıdır. Bu kural, X ürünü içeren işlemlerin, Y ürünü içermesi eğiliminde olduğunu belirtmektedir. BK algoritmaları iki önemli parametreye sahiptir: destek ve güven. Öğe kümesinin destek değeri (support), X ürünün veri tabanındaki toplan sayısının, veri tabanındaki işlem sayısına oranını göstermektedir. Güven (confidence) değeriyse $X \cup Y$ öğe kümesinin destek değerinin, $X \rightarrow Y$ kuralı için X öğe kümesinin destek değerine oranı olarak tanımlanmaktadır.

Bu araştırma, finans ve bankacılık sektörlerinde çok satın alınan ürün aileleri belirlemek ve bu ürünlere göre kuralları oluşturmak için BK algoritmalarını kullanır. Apriori, BK alanında en iyi bilinen algoritma olduğu için [7], bu araştırmada benimsenmiştir.

2.2 Lojistik Regresyon

Kategorik seçimleri açıklamak için, istatistiksel modellerin kullanımı oldukça yaygındır [8]. Bu seçimler, genel olarak nitel değerler olup, LR modelleri analiziyle yapılabilmektedir [9]. LR modeli, kolayca uygulanabilir ve anlaşılabilir olmasından dolayı, tıbbi ve sosyal bilimler alanlarında sıkça kullanılmaktadır. Bu alanlara ek olarak, LR modelleri ayrıca biyoloji, psikoloji, ekonomi ve ulaşımda da kullanılmaktadır [10]. 1977 yılından bu yana LR modelleri, pazarlamada etkin olarak kullanılmaktadır [11].

LR modelinin temel amacı, En Küçük Kareler Regresyonu, Eğri Uydurma ve Genelleştirilmiş Doğrusal Regresyonu gibi istatistikte kullanılan diğer tekniklerle aynıdır. Bu teknikler, bağımlı ve bağımsız değişkenler arasındaki ilişkiyi temsil edip, aralarındaki en uygun ilişkiyi bulmaya çalışırlar. İlişkiler $y = f(x)$ gibi basit bir fonksiyon ile temsil edilebilir. Bu fonksiyonda y ve x sırasıyla bağımlı ve bağımsız değişkenlerdir. İstatistiksel ilişkiler göz önüne alındığında, x değeri biliniyor ise y değeri belirli bir hata terimi ile tahmin edilebilir.

Aşağıdaki formül lojistik kümülatif dağılım fonksiyonunu tanımlamak için kullanılır.

$$y = f(x) = \frac{1}{1 + e^{-z}} \tag{1}$$

z, "Fayda Fonksiyonu" olarak tanımlanmakta ve aşağıdaki şekilde gösterilmektedir.

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon \tag{2}$$

Bu formüle göre β_0 sabiti, $\beta_1 x_1, \beta_2 x_2, \ldots, \beta_n x_n$'ler ise tahmin değerleriyle çarpılan regresyon katsayılarını temsil etmektedir. n sayısı veri kümesi içerisindeki bağımsız değişkenlerin sayısını, ε ise hata terimini ifade eder.

LR, iki terimli, ordinal veya çok terimli olarak sınıflandırılabilir. Eğer bağımlı değişken için gözlemlenen değerler, "Evet","Hayır" veya "Başarılı","Başarısız" gibi sadece iki tip barındırıyorsa, iki terimli LR kullanılabilir. Eğer bağımlı değişken değerleri üç veya üçten fazla terim barındırıyorsa, burada ancak çok terimli LR kullanılabilir. Tipler sıralıysa, ordinal LR kullanılabilir. Ürünlerin konfigürasyonları ve bankacılık sektöründe var olan veriden (3. bölümde detaylandırılmıştır.) dolayı, bu çalışmada müşterilerin ürün seçenekleri iki terimli LR ile modellenmiştir.

Genelde iki terimli LR'da bağımlı değişkenlerin tipini ifade etmek için "0" ve "1" kullanılır [12], ör: "1" satın alınmış, "0" satın alınmamışı ifade eder. Bir veya daha fazla sayıda, sürekli veya kategorik bağımsız değişkenlere(demografik veya sosyoekonomik gibi) dayalı olarak, iki terimli LR, bir gözlemin, bağımsız değişkenin iki tipinden hangisine, hangi olasılıkla tekabül ettiğini tahmin eder [13]

LR kesin meydana gelen bir olayın (satın alınma veya alınmama gibi) olasılığının tahmini için kullanıldığında, Maksimum Olabilirlik (Maximum Likelihood-MO) yaklaşımı, çoklu iterasyonlarla farklı çözümler deneyerek, gözlenen ve tahmin edilen değerler arasındaki olası en küçük sapmayı bulmayı esas alır. MO, gözlemlenen değerlerin elde edilme olasılığını maksimize eden regresyon katsayılarını verir [14]. Oluşturulan model ve elde edilen katsayılara göre modellenen ürüne sahip olmayan bir müşteriye, ürünün önerilip, önerilmemesinin kararı verilir.

2.3 Toplu Öğrenme

TÖ, tekli sınıflandırıcı yerine, çoklu sınıflandırıcıların sonuçlarının kombinasyonuna dayalı, daha iyi bir öngörü modeli yaratmayı amaçlamaktadır. Sınıflandırıcıların çıktılarının birleştirilmesi için kullanılan İstifleme (Stacking) Yöntemi ilk

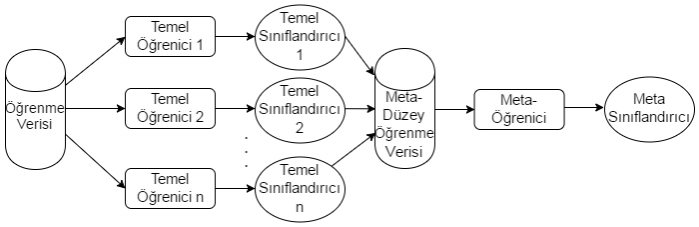
olarak David Wolpert tarafından duyurulmuştur [15]. Sonrasında Dzeroski ve Zenko'nun çalışmasına göre, İstifleme yöntemi heterojen sınıflandırıcıların birleştirilmesi için en iyi yöntem olarak bahsedilmiştir [16].

Bugüne kadar, TÖ yöntemi araştırmacılar tarafından pek çok farklı disiplinde kullanılmıştır. Bu disiplinler arasında, örüntü tanıma, istatistik ve makine öğrenmesi öne çıkmaktadır [17]. Bu yönteme göre, bu çalışmada da LR ve BK sonuçlarının birleştirilmesi amaçlanmaktadır.

Genel olarak TÖ yöntemlerinde birleştirme iki kategoride incelenir [18]. Bunlar:

- Ağırlıklandırılmış Yöntemler: Eğer sınıflandırıcılar aynı görev üzerinde çalışıyor ve sonuçlar karşılaştırılabiliyorsa kullanılması uygundur [19]. Bu alandaki en bilindik yöntemlerden biri Oylamadır (Voting).
- Meta-Düzey Yöntemler: Bu yaklaşımda, birçok temel sınıflandırıcı ve bir meta sınıflandırıcı bulunmaktadır. İlk olarak temel sınıflandırıcılar aracılığıyla öngörüler yapılır, sonrasında bu sonuçlar kullanılarak öğrenme gerçekleştirilir ve meta seviye sınıflandırıcı yaratılmış olur. Meta sınıflandırıcının sonucu, aynı zamanda sistemin sonucudur. Bu alandaki en popüler teknik İstiflemedir [16].

LR ve BK farklı ölçeklerde sayısal değerler verdiği için, bu sınıflandırıcıların sonuçları direkt olarak değerlendirilemez. Bu koşullar ve heterojen sınıflandırıcılarla çalışılması göz önüne alındığında, bu çalışma İstifleme yöntemine kaymıştır. İstifleme yöntemi, Şekil 1'de ki Philip Chan'ın çalışmasında [20] görüldüğü üzere, 2 aşamaya sahiptir. Farklı sınıflandırıcıların çıktılarını nasıl birleştirileceğiyle ilgili pek çok çalışma olmasına rağmen, LR ve BK sonuçlarının birleştirilmesi özelinde bir çalışma yoktur. İstifleme hakkında yapılan literatür araştırmasına göre, StackingC algoritmasının Meta seviyede Çok Yanıtlı Doğrusal Regresyon (ÇYDR-(Multi-response Linear Regression (MLR)) yöntemini kullandığı görülmüştür [21,22]. Bu yöntem, heterojen sınıflandırıcılar tarafından hesaplanan olasılık değerlerini birleştirme için kullandığından, bizim çalışmamıza uygundur.



Şekil 1: Meta Öğrenme

LR ve BK, her müşteri için ayrı ayrı her bir ürünün satın alınma olasılıklarını hesaplar. ÇYDR yöntemiyse, farklı algoritmalar tarafından üretilen bu olasılıkları

girdi olarak kullanılır. Bu girdi üzerinde her ürün için ayrı ayrı doğrusal regresyon çalışır ve yine her ürün için bir çıktı üretilir (Ürün satın alındıysa 1, aksi halde 0).

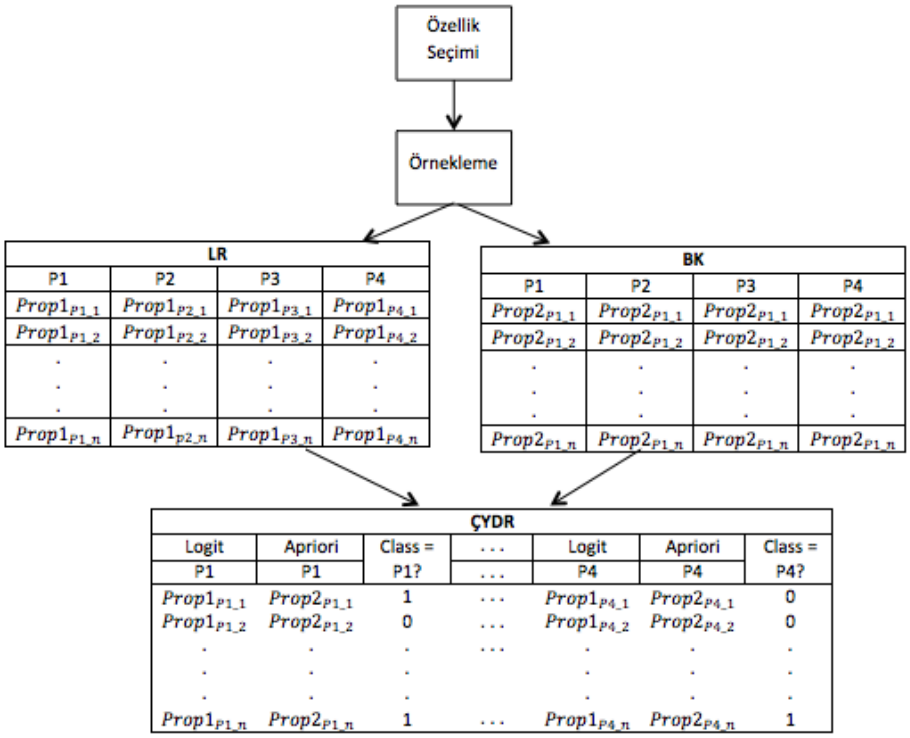
Sonuç olarak her ürün için ayrı ayrı birer model oluşturulur. Sonrasında yeni bir örnek (müşteri) için tahmin sorulduğunda, tüm ürün modelleri bu örnek üzerinde yürütülür ve ayrılan modellerin sonuçları, öğrenme aşamasında belirlenen ÇYDR katsayıları kullanılarak birleştirilir. Sonuç olarak, bir müşteri sıfır veya birden fazla ürünün alıcısı olarak sınıflandırılabilir.

İki terimli LR sonucu, her ürün için farklı olasılık değerleri ürettiği için özel düzenlemelere ihtiyaç duymaz. Bu sonuçlar ÇYDR için doğrudan kullanılabilir. Fakat, BK tarafından üretilen olasılık değerinin belirlenmesi sorunludur. İlk yaklaşım olarak, BK tarafından öngörülecek her bir ürünün olasılığı, ilgili kuralın güven değeriyle tanımlanır. Eğer bir ürün birden fazla kuralda yer alır ve belirli bir müşteri için birden fazla kural geçerli olursa, geçerli kurallar arasından en yüksek güven değerine sahip kuralın güven değeri olasılık değeri olarak belirlenir.

3 Model

Bölüm 2 de belirtildiği gibi, LR ve BK pazarlama literatüründe müşterinin bir ürünü satın alma olasılığını tahmin etmek için yaygın olarak kullanılan yaklaşımlardır. İçerisinde yaş, cinsiyet, gelir gibi müşteri demografik özellikleri ve dört farklı banka perakende ürünü için sahiplik bilgileri yer alan, Türkiye'deki orta ölçekli bir banka veri kümesi düşünüldüğünde, bu yaklaşım kabul edilebilir. Ürün sahiplik bilgileri ikili formattadır. "1" ürüne sahip olunduğunu, "0" ise sahip olunmadığını temsil eder. Bankaların her bir ana ürün ailesi için çeşitli ürünleri olduğundan, ürünler ürün ailelerine göre gruplanmış ve sahiplikler yeniden düzenlenmiştir.

Şekil 2 de gösterildiği gibi, ilk olarak eğitim veri kümesi algoritmalar için özellik seçimi (Şekil 2a) ve örneklem (Şekil 2b) adımları (4. bölümde detaylandırılacaktır) kullanılarak hazırlanmıştır. LR ile modelleme yaparken sahiplik bilgileri bağımlı, demografik bilgilerse bağımsız değişkenler olarak kullanılmıştır. Ürün aileleri birbirlerine eşdeğer olmadığından ve birbirlerini alternatif olarak temsil etmediğinden, her ürün ailesi için sonuçlar ayrı ayrı hesaplanır ve hepsi için farklı ikili doğrusal regresyon modeli oluşturulur. Aynı veri kümesi, demografik bilgiler dışında Apriori algoritmasıyla BK içinde kullanılmıştır. Her öğrenme veri kümesi için, tüm müşterilerin bütün ürün aileleri için sahiplik bilgisi kullanılarak, SRÖK oluşturulur. Ardından bu öğe kümeleri işlenerek birliktelik kuralları çıkartılmaktadır. LR modeli her ürün ailesi için ayrı ayrı uygulamak gerekirken, BK tek seferde tüm ürünler için uygulanabilmektedir. Bunun sonucu olarak iki terimli LR sonuçlarının her ürün için farklı ölçekte olması ve direk olarak karşılaştırılamaz olması nedeniyle, sonuçlar hiç bir müşteri için doğrudan birleştirilemez. Bu yüzden, LR ve BK'dan elde edilen sonuçlar, her bir ürün için ayrı ayrı ele alınmalıdır. Buda bizim LR ve BK'yı birleştirmede ÇYDR'ye ihtiyaç duymamızın nedenidir.



Şekil 2: ÇYDR Algoritmasının, LR ve BK ile Örneği

Şekil 2'de, ÇYDR algoritmasının iki baz sınıflandırıcı ile dört sınıf (P1, P2, P3, P4) ve n gözlemli bir veri kümesi üzerindeki çizimi verilmiştir. Burada $Propi_{jk}$, k. örneğin, j. ürünün, i. temel sınıflandırıcı tarafından hesaplanan olasılı- ğını göstermektedir. Şekil 2c'de LR ve BK'nın öngördükleri olasılıklar ayrı ayrı gösterilmektedir. Şekil 2d'deyse, ÇYDR için doğrusal regresyon fonksiyonu kullanılarak, her ürün için sahipliğin tahmin edileceği veri kümesi gösterilmektedir. Temelde, ÇYDR her sınıflandırıcı için doğrusal regresyon fonksiyonunu öğrenir ve sonrasında hesaplanan katsayılar, ilgili deneme veri kümesinde, sahiplikleri tahmin etmek için kullanılır.

LR, BK ve ÇYDR'nin katsayıları öğrenme aşamasının sonunda elde edilir. Bu öğrenilmiş parametreler kullanılarak, deneme veri kümesi üzerinde, her bir müşteri ve ürün için ayrı ayrı sahiplik (satın alma) olasılığı hesaplanır.

Genel olarak ÇYDR, sınıfları eşit sayıda olan çok sınıflı iki veya daha fazla sınıflandırıcıyı birleştirmek için kullanılır. Bu nedenle meta eğitim kümesinde, her sınıflandırıcı için tüm gözlemlerin olasılıklarının toplamı bire eşittir. Bu çalışmada, LR modellerinin her bir ürün için birbirlerinden bağımsız olarak olasılık oluşturulmasında herhangi bir kısıt yoktur. Çünkü ürün aileleri birbirlerine eşdeğer değillerdir.

4 Uygulama ve Test

Uygulama ve testler, yeni bir veri madenciliği aracı olan PROPCA üzerinde, Java programlama diliyle yapılmıştır. PROPCA özellik seçimi, örneklem vb. gibi gerekli tüm veri madenciliği bileşenlerini içermektedir. Ayrıca PROPCA, iki paralel ürün öneri modeli olan LR ve BK'yı da içermektedir. Bu iki modelin veri açısından farklı gereksinimleri vardır.

Önceden bahsedildiği gibi, LR modelleri, müşterilerin demografik özellikleri ve satın alma durumları arasındaki ilişkidir. Bankacılık veri kümeleri yüzlerce demografik özellik içerir. Modelleme sırasında tüm özelliklerin kullanılması, zaman tüketen ve hata eğilimli bir yaklaşım olacaktır. Bu yüzden, özellik kümesi içerisinden, modelleme işi için uygun olanlar belirlenmelidir. Bankacılık alanından sıkça başvurulan özellikleri arasından güçlü olanlar, PROPCA'nın içerisindeki bilgi değeri algoritması kullanılarak seçilir. Bu özellikler arasında hem sayısal hem de kategorik değerlere sahip olanlar bulunabilmektedir.

Nihai bankacılık veri kümesi, bilgi değeri tarafından seçilen 5 demografik özelliğe (3 sayısal ve 2 kategorik) ve 4 ürün ailesine sahip 19.818 müşteriye sahiptir. Bankacılık verisinde müşteri gizliliğinden dolayı, kullanılan özelliklerin ve ürünlerin isimlerinden bahsedilmemektedir. LR'nin uygulanabilmesi için kategorik özelliklerin kukla değerleriyle ifade edilmesi ve tüm özelliklerin standardize hale getirilmesi gerekmektedir.

Tablo 1: Sahiplik Oranı (≈)

	P1	*P2*	*P3*	*P4*
Sahiplik	%32.2	%25.3	%62.2	%63

Tablo 1'de bütün ürün aileleri ve tüm müşteriler için bu ürünlerin sahiplik oranları sunulmaktadır. Bazı müşteriler bu ürünlerden birden fazlasına sahiptir. BK'da bu ürünlerin sahiplikleri arasındaki ilişkiyi tespit ederek, yeni bir ürün satın alabilecek müşterileri algılar. Bu işlemi gerçekleştirmek için, asgari destek ve asgari güven değerlerine ihtiyaç duymaktadır. Bu değerler sırasıyla 0.01 ve 0.20 olarak tespit edilmiştir. Asgari destek değeri, SRÖK'leri bulmak için kullanılırken, asgari güven değeriyse bu öğe kümelerinden kurallar oluşturmak için kullanılır.

Kullanılan veri kümesi PROPCA'da yer alan katmanlı örnekleme kullanılarak 3 eşit parçaya ayrılmıştır. Her bir parça eğitim kümesi olarak kullanılırken, verinin geri kalanı yinelemeli olarak test için kullanılmaktadır. Bu nedenle, farklı eğitim verilerinin genel sonuçlara etkisi gözlemlenmiştir. Herbir veri kümesi, 6606 farklı müşteriye ve her ürün için farklı sahiplik oranlarına sahiptir. (Tablo 2 de görülebilir.)

Her yinelemede model, bir veri kümesiyle öğrenme gerçekleştirirken, geri kalan 2 veri kümesinin bütünüyle test etmektedir. Belirlenen bir eşik değeriyle (i.e. 0.5) sınıfların "1" veya "0" değeri elde edilir ve böylece algoritmaların tahmin

Tablo 2: Farklı Eğitim Veri Kümelerindeki Sahiplik Oranı (≈)

	P1	*P2*	*P3*	*P4*
Eğitim Kümesi 1	%32.3	%25.7	%62.1	%63.0
Eğitim Kümesi 2	%32.1	%24.8	%61.6	%63.2
Eğitim Kümesi 3	%32.1	%25.3	%63.0	%62.8

olasılık- ları değerlendirilir. Duyarlılık (Sensivity - True Positive Rate), Özgüllük (Specifity - True Negative Rate) ve Doğruluk (Accuracy) değerleri, her bir ürün ailesi için farklı test kümelerinde hesaplanır. Bu testler, farklı algoritmaların ÇYDR ile birleştirilmesinin faydalarını göstermeyi amaçlamaktadır.

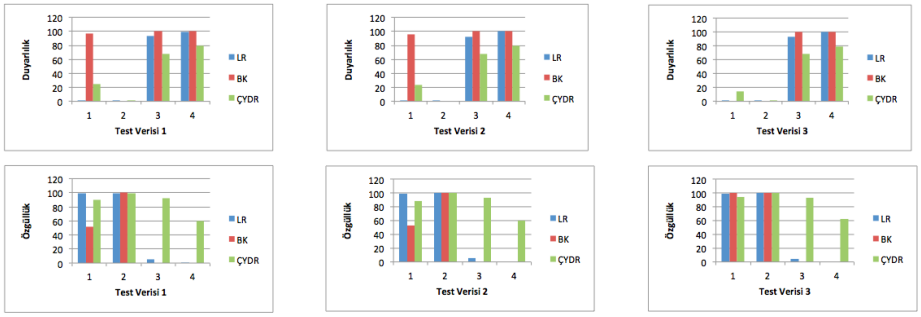
4.1 Duyarlılık ve Özgüllük

Duyarlılık ve Özgüllük birbirini tamamlayıcı göstergelerdir. Duyarlılık doğru tespit edilen pozitiflerin (e.g. ürünün satın alınması) oranını ölçerken, özgüllük doğru tespit edilen negatiflerin (e.g. ürünün satın alınmaması) oranını ölçer. Şekil 3'te farklı algoritmalar (LR, BK, ÇYDR) ile her bir test kümesinden (Test kümesi 1, 2, 3), her ürün (Product 1, 2, 3, 4) için elde edilen duyarlılık ve özgüllük değerleri gösterilmiştir. Tüm testlere birinci ürün için bakıldığında, LR'nin duyarlılık ve özgüllük değerleri yaklaşık olarak sırasıyla %0 ve %100 olarak çıkmaktadır. LR, P1'i satın alanlarla almayanları ayırt edememekte ve müşterilerin önemsiz bir kısmını alıcı olarak sınıflandırma eğilimindedir. Bu başarısız sonuçların nedeni, birinci ürünün (%50'den az) veri kümelerinde (Tablo 1 ve 2) düşük sahiplik oranlarına sahip olması olabilir. BK için 1. ve 2. test verisine bakıldığındaysa duyarlılık ve özgüllüğün sırasıyla %95 ve %50 olduğu, ancak 3. test kümesindeyse hiçbir müşteriyi alıcı olarak sınıflandırmama eğilimindedir. Bu nedenle duyarlılık %0, özgüllükse %100'dür. İlk iki test için ÇYDR sonuçları LR ve BK sonuçları arasındadır. Son testteyse ÇYDR'nin duyarlılığı (≈ %14) iki temel sınıflandırıcıdan büyükken, özgüllük (≈ %95) sonuçları küçüktür. Buradan özetle, özgüllük tarafında küçük bir kayıp olmasına rağmen, duyarlılık sonuçlarında %14 kazanç sağlanmıştır.

Ürün 2 için yok denecek kadar az müşteri, alıcı olarak doğru sınıflandırılabilmiştir. Bu nedenle, duyarlılık değerleri tüm testlerde zayıf kalmıştır. Ne LR ne de BK, ürün 2 için satın alma modeli olabilir. Bunun en büyük nedeni, 2. ürünün tüm veri kümelerinde %25 gibi kötü bir sahiplik oranı olmasıdır. Aksine, özgüllük sonuçları tüm testlerde neredeyse %100'dür. Bu durumda, ÇYDR sonuçlarda düzeltme yapamamıştır.

BK, 3. ve 4. ürün için testlerin tümünde müşterileri satın alır olarak sınıflandırmıştır. Bu yüzden, BK için duyarlılık sonuçları her zaman %100 çıkarken, özgüllük sonuçları %0'dır. BK, alan veya almayanı ayırt edememiştir. Öte yandan LR sonuçları da BK'dan çok farklı değildir. duyarlılık sonuçları %90'dan fazla iken, özgüllük sonuçları %10'dan aşağıdadır. Burada ÇYDR, duyarlılık ve özgüllük sonuçlarını dengeler. 3. ve 4. üründe farklı test kümelerinde, duyarlılıkta %10-35 arası kayıp yaşanırken, özgüllükte %60-90 arasında kazanç

sağlanmıştır. Özetle, ÇYDR iki zayıf sınıflandırıcı sonuçlarını girdi olarak alarak, duyarlılık ve özgüllük arasındaki dengeyi artırmıştır. Ayrıca, üçüncü test için 1. ürüne bakıldığında ÇYDR'nin duyarlılık sonuçlarını iyileştirdiği görülmektedir. Bir sonraki bölümde doğruluk göstergesi ve ÇYDR'nin bu göstergeye etkisi ele alınacaktır.

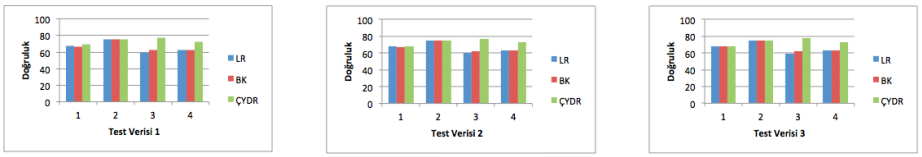


Şekil 3: Duyarlılık ve Özgüllük Sonuçları

4.2 Doğruluk

Doğruluk ölçüsü, sınıflandırıcıların pozitif ve negatif tahminlerini birlikte kullanır. Şekil 4'te görüldüğü gibi, ÇYDR, hem LR hem BK'nın doğruluklarını artırmıştır. Daha belirli olması açısından ürün 2 dışında doğrulukları tüm veri kümeleri için incelediğimizde, ürün 1 için yaklaşık %2, ürün 3 ve 4 içinse %8'den fazla ÇYDR ile kazanç elde edilmiştir.

Tüm ürün aileleri için doğruluk değerlerinin %60 civarında olmasından dolayı, kullanıcı davranışlarının karmaşık olduğu şeklinde basitçe yorumlanabilir. Özellikle müşteriye bir şey tavsiye edileceği zaman, bu daha da karmaşıklaşır. Daha önce bilinen birisine bir şey önermekle, hiç bilinmemiş ya da duyulmamış birisine öneride bulunmak aynı değildir. Bu yüzden doğruluk değerinin yüksek olmaması sosyal bilimlerde kabul edilebilir bir şeydir. Şekil 4'ün ışığında, ÇYDR'nin



Şekil 4: Doğruluk Sonuçları

neredeyse tüm test verileri için, satın alan müşterilerle, satın almayan müşterileri, LR ve BK'dan daha iyi ayırdığı söylenebilir.

5 Sonuçlar ve Gelecek Çalışmalar

Bu çalışmada, PROPCA'nın bir parçası olarak geliştirilen iki ayrı algoritmanın (LR ve BK) çıktılarının birleştirilmesi amaçlanmıştır. Fakat büyük veri kümeleri ve farklı işlem ilkeleriyle efektif çalışma konularında zorluklar bulunmaktadır. Sonuçlar umut verici olmakla birlikte, bu modelin oldukça efektif ve diğer iki modelden daha iyi doğruluğa sahip olduğu görülmüştür. Bu çalışma sonucundaki modelin, diğer iki modelin tahmin gücünü kuvvetlendirdiği ve geliştirdiği kabul edilir. Ayrıca yazarlar, modeli doğrulamak için daha fazla ürün ve müşteri bulunan bir veriyle çalışmanın genişletilmesi gerektiğinin farkındadır.

Veri kümesindeki pozitif ve negatif sınıfların dengeli olması halinde LR'nin daha iyi çalıştığı bilinmektedir. Örnekleme modelinin dikkatli olarak oluşturulması halinde modelin doğruluğunun artması mümkündür. Eğer her bir ürün için daha dengeli bir veri kümesi (yaklaşık %50 pozitif, %50 negatif gözlem) yaratılırsa, LR'nin performansı bu örnekleme kıyasla artabilir. Bunun sonucunda, LR çıktısı daha iyi öngörü olasılıkları vererek, performans artışı sağlayabilir ve bu da genel modelin başarısını olumlu etkileyebilir.

Ayrıca her ürün için ayrı ayrı, dengeli bir eğitim veri kümesinin ele alınması, farklı güçlü özelliklerin seçilmesi ve farklı kesme (cut-off) noktalarının belirlenmesi, LR modelinin ve bunun sonucunda ÇYDR modelinin başarısını artırabilir. Önerilen modelin verimliliğini artırmak için, gelecekte bazı öncelikli çalışmalar yapılabilir. BK, veri kümesine uygulamadan önce, benzer demografik özelliklere sahip müşterileri bir araya toplamak amacıyla kümeleme(ör: basit k ortalamalı, beklenti maksimizasyonu [23]) gibi ek adımlar uygulanabilir. Böylece her kümedeki müşteriler için, BK ayrı ayrı çalıştırılabilir.

Benzer şekilde, Gizli Sınıf (Latent Class) yaklaşımı LR için kullanılabilir. Gizli Sınıf analizi, sınıf sayılarının doğru seçimi için önemli bir konudur. Akaike Bilgi Kriteri [24] ve Bayesian Bilgi Kriteri [25] gibi olasılık kriterleri farklı sayıda sınıfa sahip modelleri karşılaştırmada kullanılabilir. Örneklemdeki müşteriler, müşteri özellikleri ve ürün tercihlerine göre segmentlere ayrılmış olabilir. Sonrasında her segment için karakteristik değerler, segmentte yer alan müşterilerin özellikleri kullanılarak hesaplanabilir. Evrendeki her müşterinin seçilmiş özellikleri kullanılarak, her segmentin mesafeleri hesaplanabilir, müşteriler en az uzaklığa sahip oldukları segmentlere atanabilir ve tahmin olasılıkları ilgili segment özellikleri kullanılarak hesaplanabilir.

Kaynaklar

1. Silverpop. "Silverpop email marketing metrics benchmark study.", 2014.
2. Tanguma, Jesus, and Roberto Saldivar. "Interpretation of Logistic Regression Models in Marketing Journals." The Sustainable Global Marketplace. Springer International Publishing, 2015. 2-2.
3. Akinci, Serkan, et al. "Where does the logistic regression analysis stand in marketing literature? A comparison of the Market Positioning of Prominent Marketing Journals." European Journal of Marketing 41.5/6 (2007): 537-567.

4. Brin, Sergey, Rajeev Motwani, and Craig Silverstein. "Beyond market baskets: Generalizing association rules to correlations." ACM SIGMOD Record. Vol. 26. No. 2. ACM, 1997.
5. Kotsiantis, Sotiris, and Dimitris Kanellopoulos. "Association rules mining: A recent overview." GESTS International Transactions on Computer Science and Engineering 32.1 (2006): 71-82.
6. Srikant, Ramakrishnan, and Rakesh Agrawal. Mining generalized association rules. IBM Research Division, 1995.
7. Hegland, Markus. "Algorithms for association rules." Advanced lectures on machine learning. Springer Berlin Heidelberg, 2003. 226-234.
8. Flath, David, and E. W. Leonard. "A comparison of two logit models in the analysis of qualitative marketing data." Journal of Marketing Research (1979): 533-538.
9. Berkson, Joseph. "Maximum likelihood and minimum X 2 estimates of the logistic function." Journal of the American Statistical Association 50.269 (1955): 130-162.
10. Malhotra, Naresh K. "The use of linear logit models in marketing research." Journal of Marketing research (1984): 20-31.
11. Green, Paul E., Frank J. Carmone, and David P. Wachspress. "On the analysis of qualitative data in marketing research." Journal of Marketing Research (1977): 52-59.
12. S. Hosmer, David W. Lemeshow. "Log-linear models." Springer-Verlag, 1990, ISBN:0-471-35632-8.
13. De La Viña, Lynda, and Jamie Ford. "Logistic regression analysis of cruise vacation market potential: Demographic and trip attribute perception factors." Journal of Travel Research 39.4 (2001): 406-410.
14. McCullagh, Peter, and John A. Nelder. Generalized linear models. Vol. 37. CRC press, 1989.
15. Wolpert, David H. "Stacked generalization." Neural networks 5.2 (1992): 241-259.
16. Džeroski, Saso, and Bernard Ženko. "Is combining classifiers with stacking better than selecting the best one?." Machine learning 54.3 (2004): 255-273.
17. Sewell, Martin. "Ensemble learning." RN 11.02 (2008).
18. Rokach, Lior. "Ensemble-based classifiers." Artificial Intelligence Review 33.1-2 (2010): 1-39.
19. Sigletos, Georgios, et al. "Combining information extraction systems using voting and stacked generalization." The Journal of Machine Learning Research 6 (2005): 1751-1782.
20. Fan, David W., Philip K. Chan, and Salvatore J. Stolfo. "A comparative evaluation of combiner and stacked generalization." Proceedings of AAAI-96 workshop on Integrating Multiple Learned Models. 1996.
21. Seewald, Alexander K. "How to make stacking better and faster while also taking care of an unknown weakness." Proceedings of the nineteenth international conference on machine learning. Morgan Kaufmann Publishers Inc., 2002.
22. Ting, Kai Ming, and Ian H. Witten. "Issues in stacked generalization." J. Artif. Intell. Res.(JAIR) 10 (1999): 271-289.
23. Bottou, Leon, and Yoshua Bengio. "Convergence properties of the k-means algorithms." Advances in Neural Information Processing Systems 7. 1995.
24. Akaike, Hirotugu. "Factor analysis and AIC." Psychometrika 52.3 (1987): 317-332.
25. Schwarz, Gideon. "Estimating the dimension of a model." The annals of statistics 6.2 (1978): 461-464.